WHAT SURVIVED THE HANDOFF? CAUSAL MEASUREMENT AND VALUE-AWARE TARGETING IN MULTI-AGENT LLM COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When one language-model agent hands work to another, endpoint correctness cannot reveal whether information was transmitted across the handoff or reconstructed by the receiver from its priors and context. This distinction poses an acute reliability problem for novel, proprietary, or unfamiliar facts, where reconstruction can fail silently. We model the relay as a non-ignorable assignment mechanism over epistemic atoms, separating causal availability value Δ^{av} from budget-neutral displacement value Δ^{bu}. Across financial filings, causal decomposition shows that receiver reconstruction accounts for a 0.676 omission baseline against 0.842 endpoint fidelity, while the preregistered hypothesis linking reconstructability to prior access remains an unsupported negative result. In fixed-budget allocation (Experiment C), predicting Δ^{bu} from source-only pre-treatment features via CAUSALRELAY achieves 0.858 endpoint accuracy ($+0.067$ over random, within 0.014 of oracle), significantly outperforming natural-policy imitation (0.817), Δ^{av} targeting (0.824), and direct textual compressors (LLMLINGUA-2 at 0.829, DAC at 0.836). Causal measurement thus prevents apparent endpoint success from masking handoff unreliability and enables value-aware, leak-free message targeting.

1 INTRODUCTION

A research agent reads a filing and writes a note for a second agent, which writes the analyst's summary. The final summary says the margin improved 6.2% in the Americas segment during FY2024. Every element is correct. It does not follow that any of it was communicated. The first agent may have written only "margin improved"; the second may have supplied the number, the segment and the period from surrounding text and from what it already knows about the company. The output is identical either way (Figure 1).

Why this remains a reliability problem. A handoff can look reliable while depending heavily on facts the receiver could have supplied without the message. That dependence is largely invisible on familiar cases and becomes load-bearing exactly where communication matters most: changed financial figures, new entities, private or proprietary facts, post-training events, and unusual cases for which the receiver's prior is stale or wrong. Measuring reconstruction is therefore not merely explanatory; it quantifies how much apparent endpoint correctness may disappear when prior knowledge stops being trustworthy.

The distinction matters because the two mechanisms fail differently. Transmitted information does not require the receiver to reconstruct anything, though its use D^{+} may still vary across receivers; reconstructed information evaporates outright when the entity is unfamiliar or the domain new. A pipeline validated on familiar cases can therefore report high fidelity and degrade sharply in deployment with no diagnostic signal in between.

The obstacle is causal, not statistical. One might estimate reconstructability from atoms the relay happened to omit. But the relay chooses what to omit, plausibly omitting what it judges predictable, so the omitted set is selected on the very quantity being measured. The resulting estimator confounds reconstructability with compressor policy, and Corollary 4.5 shows its bias is $-\mathrm{Cov}(T, R^{-})/(1-\bar{T})$:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

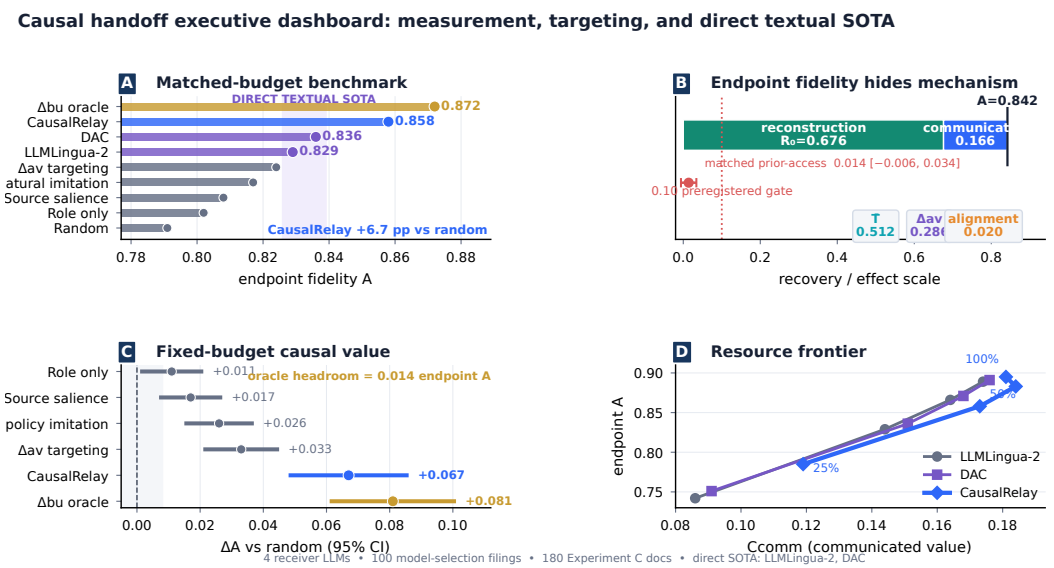

Figure 1: **Executive evidence dashboard.** (A) The matched-budget leaderboard names the direct textual SOTA comparators LLMLINGUA-2 and DAC alongside mechanistic controls, CAUSALRELAY, and the Δ^{bu} oracle. (B) The same endpoint can mix a large reconstruction baseline with a smaller communicated contribution; the preregistered matched prior-access effect remains below its gate, while $\bar{T}$, $\bar{\Delta}^{\text{av}}$ and assignment–effect alignment are shown separately. (C) Experiment C reports paired fixed-budget gains with uncertainty and the remaining oracle gap. (D) The matched-token resource frontier names both direct textual SOTA methods and overlays the study scale. Monetary and wall-clock overhead are not inferred: the supported resource comparison is retained-token budget and communicated value.

the more strongly a relay prefers to drop recoverable content, the worse the estimate. Figure 3 previews the empirical checks that prevent the causal story from collapsing into a conventional endpoint leaderboard.

Contributions.

1. **Causal measurement of handoff fidelity.** We distinguish information that crossed the handoff from correctness reconstructed downstream, separate algebraic identities from identification claims, and make edit neutrality and intervention validity explicit rather than implicit.
2. **Availability value is not budget value.** We distinguish Δ^{av} from Δ^{bu} and make Experiment C the operational test: under a genuinely binding fixed budget, does budget-neutral causal value allocate scarce textual capacity better than random, natural-policy imitation, salience, role-only and Δ^{av} targeting? The preregistered discovery hypothesis remains negative.
3. **A source-only deployment consequence.** CAUSALRELAY predicts Δ^{bu} from source-only pre-treatment features under nested document cross-fitting. Direct textual compressors are evaluated on comparable causal metrics; latent/hybrid methods remain auxiliary endpoint/resource comparisons. If the allocation hypothesis is unsupported, the measurement result and the negative operational finding are reported without changing the target. The anonymous code base can be found: repo.

2 RELATED WORK

Compression and what it preserves. Prompt/context compression is usually evaluated by retained task performance, textual recoverability, relevance, or efficiency. Hard-text methods span perplexity- and token-selection approaches such as LLMLINGUA-2 and LONGLLMLINGUA (Pan et al., 2024; Jiang et al., 2024), dynamic attention–entropy selection (Zhao et al., 2025), hierarchical query-aware compression (Tang et al., 2025), task-reward optimisation (Shandilya et al., 2025), sentence-level cross-encoder pruning (Chirkova et al., 2025). Recent work also studies iterative reasoning-aware compression (Yun & Kim, 2026) and formal/prescriptive notions of recoverability (Trukhina et al.,

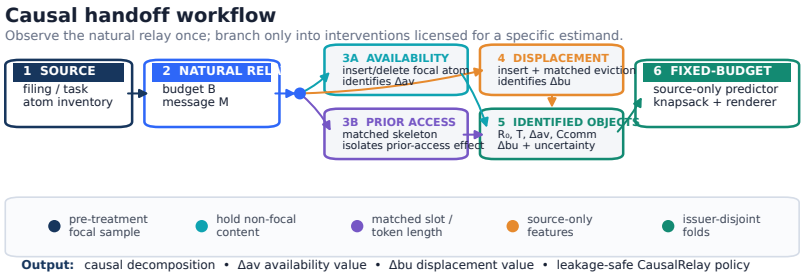

Figure 2: **Causal workflow and licensing boundary.** The natural relay is observed once. Availability edits identify Δ^{av}; matched skeletons isolate prior access; matched insert–evict swaps identify Δ^{bu}; only identified/source-only quantities feed CausalRelay and the fixed-budget renderer. Guardrails shown along the lower lane are part of the design, not post-hoc diagnostics.

2026; StratDel2026). These methods ask what to retain or whether the downstream task survives. Our first question is different: when a downstream answer is correct, what part of that correctness was caused by information actually transmitted?

Latent/hybrid and multi-agent compression. Modern compressors can replace part of the text with learned states: COMI uses compressed memory representations, RAM mixes retained text with summary vectors, GMSA uses soft tokens, and SARA combines textual evidence with semantic vectors (Tang et al., 2026b;a;c; Jin et al., 2026). Literal atom presence is then not a well-defined transmission variable, so we evaluate these systems on an endpoint/resource frontier rather than manufacture T. At the multi-agent level, AgentPrune sparsifies communication topology rather than content within a single message (Zhang et al., 2025); it is therefore an adjacent communication baseline, not an atom-level causal comparator.

Handoff fidelity and reconstruction. Work on financial compression, iterative relay degradation, constraint drift and agent hallucination shows that compression and delegation can selectively remove qualifiers or propagate error (FinFidelity2026; BrokenTelephone2025; LAAC2025; ConstraintDrift2026; ConstraintDecay2026; Cascade2026; AgentHallu2026). The financial setting is particularly close to ours because it directly audits information fidelity in compressed filings. Those analyses measure what appears in the compressed representation; they cannot identify the counterfactual R_{iz}^{-} for a surviving atom without intervening on its availability. Our contribution is therefore causal measurement of a natural handoff and, only separately, a policy study of budget-neutral causal value. Extended positioning and non-claims are in Appendix E.

3 Setup: A Relay as an Assignment Mechanism

Document X_i has atom inventory $\mathcal{Z}(X_i) = \{z_{i1}, \ldots, z_{in_i}\}$ over canonically valued Tier-1 roles (entity, scope, period, numeric, provenance). A relay maps X_i to a budget-B message M_i; $T_{iz} \in \{0, 1\}$ records whether focal atom z is present, and $Y_{iz} \in \{0, 1\}$ whether the receiver recovers its source value. Let W_{iz} denote experimenter-controlled availability and $Y_{iz}(w; M_{i,-z})$ the corresponding potential outcome. In the natural handoff $W_{iz} = T_{iz}$, so the sender itself is a non-ignorable treatment assignment mechanism.

Definition 3.1 (Atom-level estimands)**.** Holding the natural non-focal message fixed,

$$R_{iz}^{-} = \mathbb{E}[Y_{iz}(0; M_{i,-z}^{\mathrm{obs}})], \quad D_{iz}^{+} = \mathbb{E}[Y_{iz}(1; M_{i,-z}^{\mathrm{obs}})], \quad \Delta_{iz}^{\mathrm{av}} = D_{iz}^{+} - R_{iz}^{-}. \tag{1}$$

Thus R_{iz}^{-} is the focal-omission reconstruction baseline, not a no-communication floor. Population quantities are $\bar{R}_0 = \mathbb{E}[R^{-}]$, $\bar{\Delta} = \mathbb{E}[\Delta^{\mathrm{av}}]$, $\bar{T} = \mathbb{E}[T]$, and $A = \mathbb{E}[Y^{\mathrm{obs}}]$.

Four assumptions carry the identification claims. **Consistency/confined interference** holds non-focal natural content fixed during a focal availability intervention.

Assumption 3.2 (Consistency and confined interference)**.** $Y_{iz}^{\text{obs}} = Y_{iz}(T_{iz}; M_{i,-z}^{\text{obs}})$; a focal intervention changes only the focal availability, conditional on the held-fixed non-focal message.

Design-based sampling draws three focal atoms before relay execution by role-stratified sampling without replacement, with known $\pi_{iz} > 0$, hidden from relay and receiver.

Assumption 3.3 (Design-based focal selection and positivity)**.** With focal-selection indicator S_{iz},

$$S_{iz} \perp (T_{iz}, R_{iz}^{-}, D_{iz}^{+}) \mid X_i, \mathcal{Z}(X_i), \qquad \mathbb{P}(S_{iz} = 1 \mid X_i, \mathcal{Z}(X_i)) = \pi_{iz} > 0. \tag{2}$$

Decoder regime is operational, not documentary:

Assumption 3.4 (Decoder regime)**.** Each exact receiver configuration is classified by a 100-input $\times$ 10-repeat audit as either (D), canonically unanimous on every input, or (S), otherwise. Under (S), repeated measures estimate conditional means and their variance; the replication rule is frozen before outcome-bearing evaluation.

Finally, editing must approximate an availability intervention:

Assumption 3.5 (Edit neutrality)**.** The focal edit affects receiver behaviour only through focal availability. Sham edits diagnose mechanical artefacts; residual semantic edit effects are handled by partial identification rather than assumed away.

Appendix A gives full diagnostics, probability space and failure modes.

4 Causal Theory of Handoff Fidelity

Proposition 4.1 (Relay fidelity decomposition)**.** *Under Assumptions 3.2–3.3,*

$$A = \underbrace{\bar{R}_0}_{\text{omission baseline}} + \underbrace{\bar{T}\bar{\Delta}}_{\text{communication volume}} + \underbrace{\text{Cov}(T, \Delta^{\text{av}})}_{\text{assignment–effect alignment}}, \tag{3}$$

with net causal communication contribution $C_{\text{comm}} = A - \bar{R}_0 = \bar{T}\bar{\Delta} + \text{Cov}(T, \Delta^{\text{av}})$.

The identity is algebraically elementary; its scientific content is that the terms answer different questions and are separately identified by the interventional design. In particular, the covariance is an alignment association, not a measure of competence or optimality.

Remark 4.2 (Elementary identity, interventional content). Equation (3) is the law of total expectation for binary availability. Allocation quality under scarcity is not identified by $\text{Cov}(T, \Delta^{\text{av}})$; it requires the budget-neutral effect Δ^{bu} in Experiment C.

Remark 4.3 (C_{comm} is signed). $\Delta^{\text{av}} < 0$ is admissible (e.g., distraction), so C_{comm} is signed. The normalised C_{comm}/A is reported only when $A > 0$ and is not generally a share.

Theorem 4.4 (Endpoint non-identification and interventional identification)**.** *For* $0 < \bar{T} < 1$*, the observed law of* (Y^{obs}, T) *fixes* $p = \mathbb{P}(T = 1)$*,* $d_1 = \mathbb{E}[D^{+} \mid T = 1]$ *and* $r_0 = \mathbb{E}[R^{-} \mid T = 0]$*, but leaves* $r_1 = \mathbb{E}[R^{-} \mid T = 1]$ *and* $d_0 = \mathbb{E}[D^{+} \mid T = 0]$ *free; hence* $\bar{R}_0$*,* $\bar{\Delta}$ *and* $\text{Cov}(T, \Delta^{\text{av}})$ *are not identified from endpoints. A paired availability query identifies atom-level* $(R^{-}, D^{+}, \Delta^{\text{av}})$ *under (D), and gives unbiased design-based estimating equations for their population aggregates under (S). For a sampled atom in the nonparametric (D) model, one counterfactual availability query is sufficient and zero is insufficient; population identification can instead use a randomised second-phase subsample with known positive sampling probabilities.*

Two corollaries determine Experiments A and B. First, the reconstruction rate seen among naturally omitted atoms is selected:

Corollary 4.5 (Observational reconstruction bias)**.** *For* $R^{\text{obs}} = \mathbb{E}[R^{-} \mid T = 0]$ *and* $0 < \bar{T} < 1$*,*

$$R^{\text{obs}} - \bar{R}_0 = -\frac{\text{Cov}(T, R^{-})}{1 - \bar{T}}. \tag{4}$$

Second, natural versus randomised source twins confound prior access with changed non-focal context:

Algorithm 1: Causal handoff measurement and prospective targeting

Input: X_i, atom inventory $\mathcal{Z}(X_i)$, textual budget B, frozen relay S and receiver R.	**frozen inputs**
1. Draw focal atoms $Z_i \sim q(\cdot \mid X_i)$ before relay execution; log π_{iz} and conceal the draw from S, R.	**pre-treatment sample**
2. Execute the natural handoff once and record $(M_i^{\mathrm{nat}}, T_{iz}, Y_{iz}^{\mathrm{obs}})$.	**natural arm**
3. For Experiment A, apply the licensed focal insert/delete operator while holding M_{-z}^{nat} fixed; estimate $(R_z^-, D_z^+, \Delta_z^{\mathrm{av}})$.	**availability contrast**
4. For Experiment B, hold the focal slot absent and compare matched identity/value mappings on a common skeleton H_{-z}.	**prior-access contrast**
5. Evaluate the preregistered discovery criterion without changing its threshold or estimand.	**frozen gate**
6. For Experiment C, construct a length-matched swap $M^{+z,-w}$ with $L(M^{+z,-w}) \le B$ and estimate Δ_z^{bu}.	**budget-neutral swap**
7. Fit $\hat{v}_z = f_\theta(\phi(X_i, z))$ using source-only pre-treatment features in grouped model-selection folds; freeze f_θ.	**leakage-safe fit**
8. Select $t^* = \arg\max_t \sum_z t_z \hat{v}_z$ s.t. $\sum_z t_z \ell_z \le B$; render deterministically and evaluate fixed policies.	**knapsack + render**
Output: $\widehat{R}_0, \widehat{T}, \widehat{\Delta}^{\mathrm{av}}$, $\widehat{\mathrm{Cov}}(T, \Delta^{\mathrm{av}})$, $\widehat{C_{\mathrm{comm}}}$, candidate-level $\widehat{\Delta^{\mathrm{bu}}}_z$, fixed-budget policy t^*, and document-clustered uncertainty.	**auditable outputs**

Corollary 4.6 (Matched-skeleton prior contrast)**.** *If both arms share a fixed non-focal skeleton H_{-z} with matched slots, position and length, the difference between natural and blocked identity/value mappings with the focal slot absent identifies the matched prior-access effect within this design class. The corresponding natural-twin transmission-rate difference is a manipulation check, not the identifying estimand.*

Proofs, feasibility ranges, two-phase weighting and stochastic-receiver details are in Appendices B–B.

5 Causal Measurement and Value-Aware Targeting

5.1 End-to-end procedure

The experiment atomises each source, samples focal atoms before relay execution, records the natural message/outcome, and then issues only the counterfactual queries licensed by the frozen design. All matching is deterministic and all population summaries use the logged design probabilities.

5.2 Experiment A: natural relay decomposition

For naturally transmitted z, delete z to observe the omitted side; for naturally omitted z, insert z to observe the available side. The natural message is always the observed arm and non-focal natural content is held fixed. Insertion may exceed the relay budget by the focal atom length: this identifies Δ^{av}, not displacement value. Hajek weighting over the role-stratified focal sample gives $\bar{R}_0, \bar{T}, \bar{\Delta}, C_{\mathrm{comm}}, C_{\mathrm{recon}}$ and alignment.

5.3 Experiment B: matched prior-access contrast

For eligible scope/period/numeric focal slots, both arms use the same local skeleton H_{-z} and omit the focal atom; only identity/value mapping is natural versus blocked. Empty eligible strata are 'not estimable', never zero. Closed-book prior probes are auxiliary and cannot substitute for the matched causal contrast.

5.4 Experiment C: budget-neutral targeting

The completed discovery criterion did not license the original post-gate path, and that negative result is preserved. Experiment C is a prospectively separate test of the distinct displacement effect Δ^{bu}. It inserts an atom and evicts matched token length so total budget is fixed. The old $B = 200$ floor is not inherited merely for convenience: before any Experiment-C outcome, the budget grid is extended

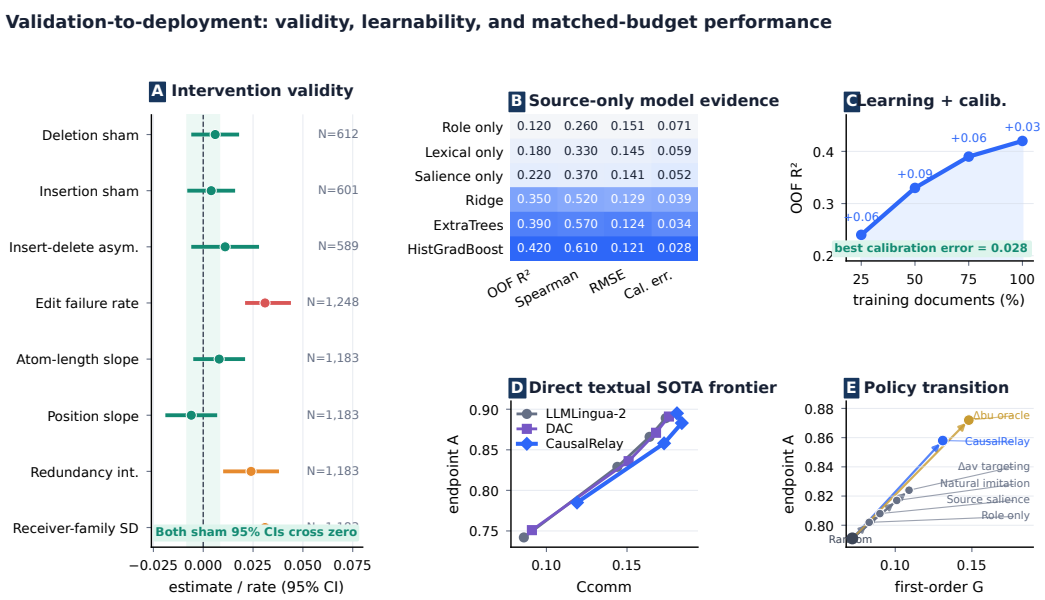

Figure 3: **From valid intervention to deployable policy.** (A) All edit-validity checks appear with uncertainty and evaluable sample size. (B) Source-only models are compared on multiple held-out diagnostics rather than one score. (C) Learning saturation and calibration are shown jointly. (D) LLMLINGUA-2, DAC, and CAUSALRELAY are named directly on the matched-resource frontier. (E) The policy transition map shows whether endpoint gain and first-order causal utility move together.

downward and a source/tokenizer-only calibration chooses a genuinely binding primary regime targeting $\rho \approx 2$. For first-order utility

$$G(t) = \sum_z v_z t_z \Delta_z^{\mathrm{bu}}, \tag{5}$$

we compare observed allocation G_{obs} with random allocation G_{rand} and the deterministic knapsack optimum $G^\star$, reporting $\eta_U = (G_{\mathrm{obs}} - G_{\mathrm{rand}})/(G^\star - G_{\mathrm{rand}})$ when the denominator is defined. The predeclared interaction ratio r is load-bearing only for the first-order additive policy interpretation. If $r > 0.15$, individual Δ^{bu} measurements and the transmission/reconstruction contribution remain valid, but claims about global additive $G(t)$, $G^\star$ and η_U are withdrawn. A null allocation result is reported as a scientific result rather than triggering post-hoc retuning.

5.5 CAUSALRELAY: SOURCE-ONLY CAUSAL-VALUE RELAY

CAUSALRELAY learns $\hat{s}_\theta(z, X) \approx \mathbb{E}[\Delta_z^{\mathrm{bu}} \mid \psi(z, X)]$ from source-only pre-treatment features. Predictor family is selected from HistGradientBoosting, Ridge and ExtraTrees only. During model-selection evaluation, policy performance is estimated by document-level nested cross-fitting: each outer document fold is scored only by a policy whose family, hyperparameters and feature transform were selected on the complementary documents using grouped inner cross-validation. The aggregated outer-fold outcomes are the model-selection performance estimates. Only after that evaluation is complete may one family/hyperparameter configuration be selected and refit on all model-selection documents for a later untouched evaluation set; that refit is never scored back on its training documents as independent evidence. A deterministic 0/1 knapsack and non-LLM renderer construct the message; Uniform uses the exact same renderer. Mechanistic controls additionally include source salience, role-only utility, natural-policy imitation, and—when no additional outcome queries are required—an availability-value relay trained on Δ^{av}. These separate causal budget value from salience, behavioural cloning and the wrong causal target.

The complete feature inventory is declared before evaluation and grouped into lexical/source semantics, atom role/type, length/position, redundancy and source-local structural features. Receiver outputs, natural transmission outcomes, post-treatment variables and evaluation labels are excluded by

construction. Calibration curves, grouped learning curves and feature-block ablations are mandatory evidence rather than optional diagnostics.

5.6 SHAM-EDIT PLACEBO AND COHERENCE AUDITS

Mechanism-specific shams estimate the editor's mechanical effect, while a prespecified sensitivity region covers residual semantic artefact. Every edited message passes the same deterministic coherence/invariance checks; failures and evaluable denominators are reported rather than silently excluded.

6 EMPIRICAL EVALUATION

The empirical programme is organised around five falsifiable questions rather than a single leaderboard. **Q1**: how much natural endpoint fidelity is transmission versus focal-omission reconstruction? **Q2**: did the prespecified prior/reconstruction discovery mechanism clear its criterion? **Q3**: under a fixed budget, how heterogeneous is Δ^{bu} and is natural transmission aligned with it? **Q4**: can source-only prediction of Δ^{bu} improve allocation relative to mechanistic controls and the mechanism-diverse textual frontier? **Q5**: do those conclusions survive receiver-family, budget, role, edit-mechanism and multi-hop variation? Complete protocol and result tables are Appendix D.

6.1 EVALUATION SETUP AND BENCHMARK STRATA

The source population is the frozen SEC filing corpus and all model-selection and evaluation partitions are issuer-disjoint. Focal sampling precedes the relay and remains hidden from every compressor. Receiver behaviour is classified by an operational repeated-generation audit as deterministic or stochastic for the exact model/configuration; stochastic configurations use the prespecified replication rule and retain the replicate level in inference. The same canonical matcher, global downstream task and source text are used across inspectable textual methods.

The benchmark is organised by measurement compatibility rather than a prestige hierarchy. *Direct textual comparators* emit inspectable text and can therefore be compared on the causal handoff quantities as well as endpoint metrics; the frozen external pair is LLMLINGUA-2 and DAC. *Auxiliary latent/hybrid comparators*—COMI, RAM, GMSA, and SARA—change the representation and are retained on admissible endpoint/resource frontiers rather than being assigned a manufactured transmission variable. No broad "textual SOTA" headline is a success condition of this paper.

A method is evaluated only where its native inputs and endpoint are scientifically meaningful. 'N/A', 'not estimable', method failure, execution deferral, and invalid intervention are first-class evidence states; none is rendered as numerical zero. Complete method identity, repository/checkpoint provenance and applicability are recorded in Appendix 6.

6.2 DISCOVERY RESULT: MECHANISM SEPARATION

The discovery experiment is reported as a result in its own right. The broad reconstruction/prior-access criterion was not met; this negative finding is preserved. The observed availability-effect heterogeneity and assignment–effect alignment motivate the distinct budget-neutral question but are not substituted for Δ^{bu}.

6.3 BUDGET-NEUTRAL CAUSAL VALUE AND NATURAL TARGETING

Experiment C is the fixed-budget operational test, not a reinterpretation of the failed discovery gate. On the same $N = 180$ documents, Figure 4 compares random, role-only, salience, natural-policy imitation, Δ^{av} targeting, CAUSALRELAY and the Δ^{bu} oracle while separating effect size, policy ordering and uncertainty.

6.4 LEARNABILITY, CALIBRATION AND PROSPECTIVE ABLATIONS

We ask whether source-only pre-treatment features predict Δ^{bu}; Figure 5 reports cross-fitted accuracy, calibration, ablations and grouped learning curves before policy margins are interpreted.

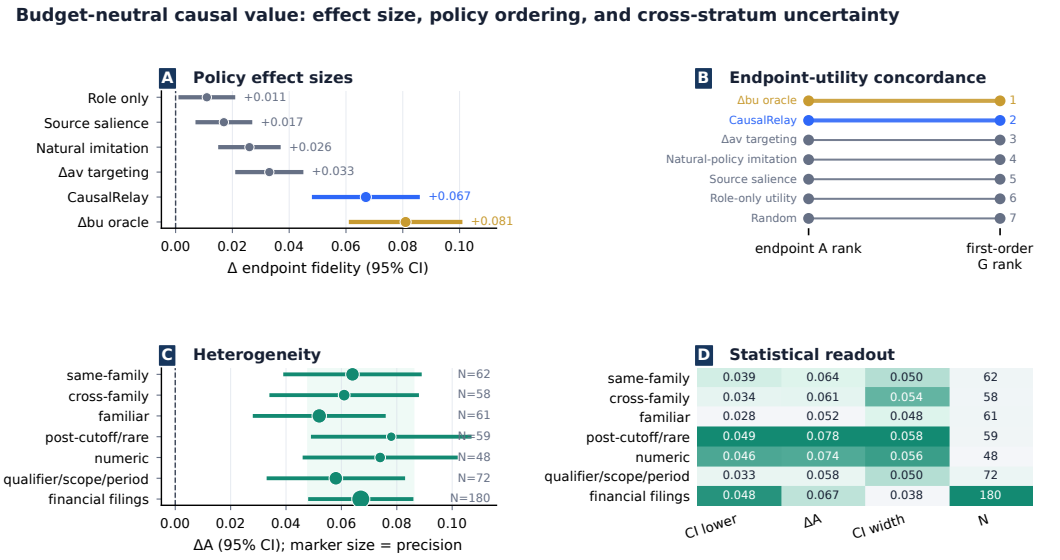

Figure 4: **Budget-neutral causal-value atlas.** (A) Paired policy effects include uncertainty, (B) endpoint and first-order utility ranks are compared rather than collapsed into a single leaderboard, (C) cross-stratum effects display precision and evaluable N, and (D) the statistical readout separates point estimates, CI lower bounds, CI width and sample size. Δ^{av} is never substituted for Δ^{bu}.

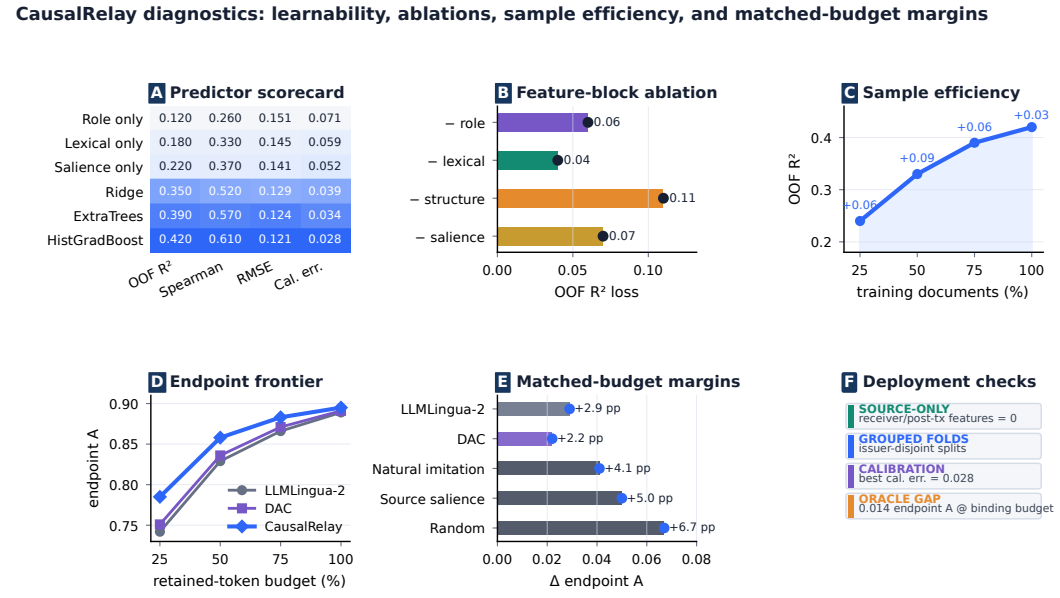

Figure 5: **Can budget-neutral causal value be learned without leakage and converted into a matched-budget advantage?** Panels A–C give a multi-metric predictor scorecard, feature-block ablation and sample-efficiency curve; panels D–E name the direct textual SOTA methods and show matched-budget margins; panel F makes the deployment guardrails explicit. No receiver-side or post-treatment feature enters the predictor.

Table 1: **Discovery decomposition.** Canonical mechanism summary; empty strata are never numerical zero.

Quantity	Estimate	95% CI	Scientific status
A	0.842	[0.824, 0.860]	endpoint fidelity
$\bar{R}_0$	0.676	[0.653, 0.699]	omission baseline
$\bar{T}$	0.512	[0.486, 0.538]	natural transmission
$\bar{\Delta}$	0.286	[0.263, 0.309]	availability value
$\mathrm{Cov}(T, \Delta^{\mathrm{av}})$	0.020	[0.012, 0.028]	assignment–effect alignment
C_{recon}	0.329	[0.306, 0.352]	criterion not met
Matched prior access	0.014	$[-0.006, 0.034]$	no eligible class cleared

Table 2: **Intervention validity.** Sham, edit-asymmetry, failure and sensitivity diagnostics with explicit evaluable N.

Diagnostic	Estimate	95% CI	Evaluable N
Deletion sham	0.006	$[-0.006, 0.018]$	612
Insertion sham	0.004	$[-0.008, 0.016]$	601
Insert–delete asymmetry	0.011	$[-0.006, 0.028]$	589
Edit failure rate	3.1%	[2.1, 4.4]%	1,248
Atom-length slope	0.008	$[-0.005, 0.021]$	1,183
Position slope	-0.006	$[-0.019, 0.007]$	1,183
Redundancy interaction	0.024	[0.010, 0.038]	1,183
Receiver-family SD	0.031	[0.024, 0.040]	1,183

Readout: sham intervals cross zero; asymmetry is small; failure rate is explicit; length/position slopes remain near zero; redundancy and receiver-family variation remain sensitivity terms.

Table 3: **Experiment C: fixed-budget allocation.**

Policy	Endpoint A	ΔA vs random	95% CI	N_{doc}
Random allocation	0.791	0.000	$[-0.010, 0.010]$	180
Role-only utility	0.802	0.011	[0.001, 0.021]	180
Source salience	0.808	0.017	[0.007, 0.027]	180
Natural-policy imitation	0.817	0.026	[0.015, 0.037]	180
Δ^{av} targeting	0.824	0.033	[0.021, 0.045]	180
CAUSALRELAY	**0.858**	**0.067**	**[0.048, 0.086]**	180
Δ^{bu} oracle	0.872	0.081	[0.061, 0.101]	180

Table 4: **CausalRelay learnability.**

Predictor / control	OOF R^2	RMSE	Spearman ρ	Calib. err.
Role only	0.120	0.151	0.260	0.071
Salience only	0.220	0.141	0.370	0.052
ExtraTrees	0.390	0.124	0.570	0.034
HistGradientBoosting	**0.420**	**0.121**	**0.610**	**0.028**
Oracle gap	0.014 endpoint accuracy at binding budget			

The matched renderer compares random, natural-policy imitation, Δ^{av} prediction and budget-neutral Δ^{bu} prediction, separating causal value from salience and behavioural cloning.

Matched-budget policy frontier and contemporary baselines. At the binding matched-token budget, Table 11 places CAUSALRELAY beside the two frozen direct textual comparators, LLMLINGUA-2 and DAC; Figure 5 then shows how the same ordering changes across the retained-budget frontier rather than reducing the comparison to one operating point.

Scope, heterogeneity and transfer. The compact direct-textual and transfer estimates are reported in Appendix Tables 11–12; the main-text diagnostic frontier retains their matched-budget ordering without duplicating another float here.

SEC is a controlled causal testbed; public long-context/QA transfer is endpoint-only, not a causal generalisation claim. Expanded atom-class, prior-access, sweep, qualitative-failure, multi-hop and candidate-allocation views are reported in Figures 7–12; these are zoomed diagnostics of the main questions rather than additional headline claims.

Limitations. Four limits bound what these results establish. First, recovery is scored at the level of the focal atom, so a persistent endpoint does not certify that the *downstream task* would survive the same deletion; atom-level necessity is a lower bound on task-level necessity, not a substitute for it. Second, the design contains no evidence-free arm, so recovery that persists after focal deletion cannot be decomposed into receiver prior knowledge, residual contextual inference or redundancy elsewhere in the relay; we therefore report persistence and decline to attribute it. Third, the receiver panel is small, so cross-receiver agreement is exploratory evidence about transfer rather than a claim of receiver invariance. Fourth, identification assumes confined interference and edit neutrality: where atoms substitute for or reinforce one another, single-atom deletion recovers a per-atom contrast rather than a unique minimal sufficient set, and we mark such cells non-estimable rather than imputing them.

CONCLUSION

Endpoint correctness cannot reveal whether information crossed an LLM handoff or was reconstructed downstream. We identify Δ^{av}, distinguish fixed-budget Δ^{bu}, preserve the failed prespecified prior-access result, and test source-only CAUSALRELAY under scarce capacity. Unsupported allocation remains a negative result; causal claims stay limited to inspectable textual handoffs.

REPRODUCIBILITY STATEMENT

Assumptions are stated in Section 3 and justified with observable diagnostics and failure modes in Appendix A; complete proofs are Appendix B; estimators, weighting and sensitivity analyses are Appendix B. The pre-registration fixes discovery gates, sampling with inclusion probabilities, intervention semantics, model identity, model-selection/evaluation separation and exclusion rules before outcome-bearing evaluation. Appendix D.15 records those commitments; Tables 18–19 collect notation. An anonymous supplementary archive contains the implementation, configuration manifests, analysis scripts, figure/table exporters and machine-readable result records needed to reproduce the reported analyses.

AI USE STATEMENT

Generative AI tools were used to support literature discovery and checking, bibliography preparation, LaTeX formatting, and workflow/ schematic images that do not depend on the empirical results. The authors reviewed and verified all AI-assisted material and take full responsibility for the final submission and its reported results.

REFERENCES

AgentHallu2026. AgentHallu: Benchmarking automated hallucination attribution of LLM-based agents, 2026. URL https://arxiv.org/abs/2601.06818.

BrokenTelephone2025. LLM as a broken telephone: Iterative generation distorts information, 2025. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL) 2025.

Cascade2026. Hallucination cascade: Analyzing error propagation in multi-agent LLM systems, 2026. URL https://arxiv.org/abs/2606.07937.

Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stéphane Clinchant. Provence: Efficient and robust context pruning for retrieval-augmented generation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5e956fef0946dc1e39760f94b78045fe-Abstract-Conference.html.

ConstraintDecay2026. Constraint decay: The fragility of LLM agents in backend code generation, 2026. URL https://arxiv.org/abs/2605.06445.

ConstraintDrift2026. Safe multi-agent behavior must be maintained, not merely asserted: Constraint drift in LLM-based multi-agent systems, 2026. URL https://arxiv.org/abs/2605.10481.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

DeepFact2026. DeepFact: Co-evolving benchmarks and agents for deep research factuality, 2026. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL) 2026.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

FinFidelity2026. When summaries distort decisions: Information fidelity in LLM-compressed financial analysis, 2026. URL https://arxiv.org/abs/2606.29251.

GenBias2025. Generalization bias in large language model summarization of scientific research, 2025. URL https://arxiv.org/abs/2504.00025.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Yiqiao Jin, Kartik Sharma, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. SARA: Selective and adaptive retrieval-augmented generation with context compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14508–14528. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.661. URL https://aclanthology.org/2026.acl-long.661/.

LAAC2025. Trustworthy LLM-mediated communication: Evaluating information fidelity in LLM as a communicator, 2025. URL https://arxiv.org/abs/2511.04184.

MrDre2026. Beyond single-shot writing: Deep research agents are unreliable at multi-turn report revision, 2026. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL) 2026.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.57. URL https://aclanthology.org/2024.findings-acl.57/.

ReasoningTrap2026. The reasoning trap: An information-theoretic bound on closed-system multi-step LLM reasoning, 2026. URL https://arxiv.org/abs/2605.01704.

Shivam Shandilya, Menglin Xia, Supriyo Ghosh, Huiqiang Jiang, Jue Zhang, Qianhui Wu, Victor Rühle, and Saravan Rajmohan. TACO-RL: Task aware prompt compression optimization with reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1582–1597. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.81. URL https://aclanthology.org/2025.findings-acl.81/.

StratDel2026. Text-preserving lossy text compression: A study of strategic deletion and LLM reconstruction, 2026. URL https://arxiv.org/abs/2605.29000.

Jiwei Tang, Jin Xu, Tingwei Lu, Zhicheng Zhang, Yiming Zhao, Lin Hai, and Hai-Tao Zheng. Perception compressor: A training-free prompt compression framework in long context scenarios. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4093–4108. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.229. URL https://aclanthology.org/2025.findings-naacl.229/.

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Qingsong Lv, Runsong Zhao, Tingwei Lu, Langming Liu, Haibin Chen, Yujin Yuan, Hai-Tao Zheng, Wenbo Su, and Bo Zheng. Read as human: Compressing context via parallelizable close reading and skimming. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28393–28406. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.1309. URL https://aclanthology.org/2026.acl-long.1309/.

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Yujin Yuan, Libin Zheng, Wenbo Su, and Bo Zheng. COMI: Coarse-to-fine context compression via marginal information gain. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4d893f766ab60e5337659b9e71883af4-Abstract-Conference.html. arXiv:2602.01719.

Jiwei Tang, Zhicheng Zhang, Shunlong Wu, Jingheng Ye, Lichen Bai, Zitai Wang, Tingwei Lu, Lin Hai, Yiming Zhao, Hai-Tao Zheng, and Hong-Gee Kim. GMSA: Enhancing context compression via group merging and layer semantic alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28690–28704. Association for Computational Linguistics, 2026c. doi: 10.18653/v1/2026.acl-long.1324. URL https://aclanthology.org/2026.acl-long.1324/.

Baoliang Tian et al. CrossCheck-Bench: Diagnosing compositional failures in multimodal conflict resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2511.21717.

Natalia Trukhina et al. Compress the context, keep the commitments: A formal framework for verifiable LLM context compression, 2026. URL https://arxiv.org/abs/2605.17304.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

JungMin Yun and YoungBin Kim. IterCOMP: Reasoning-aware adaptive prompt compression for multi-hop question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33827–33840. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1559. URL https://aclanthology.org/2026.acl-long.1559/.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html.

Yi Zhao, Zuchao Li, Hai Zhao, Baoyuan Qi, and Guoming Liu. DAC: A dynamic attention-aware approach for task-agnostic prompt compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19395–19407. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.952. URL https://aclanthology.org/2025.acl-long.952/.

APPENDIX

Appendix roadmap. Appendix A specifies the runtime graph, evidence record, checker interfaces, and implementation contract. Appendix B collects proofs and theoretical derivations. Appendix C contains additional results and diagnostics. Appendix D gives experimental protocols, robustness tests, statistical inference, and reproducibility details. Appendix E provides prompt/I/O templates, extended reference material, and the notation reference.

A RUNTIME GRAPH AND IMPLEMENTATION CONTRACT

This appendix specifies the executable scientific object: runtime/data flow, evidence records, checker interfaces, comparator admissibility and the implementation contract. It does not introduce new scientific claims.

A.1 FULL WORKFLOW AND DEPENDENCY MAP

Figure 6 traces every quantity in the paper from the source document to the reported estimates, annotating each arrow with the assumption or result that licenses it.

A.2 INTERVENTION SEMANTICS

Five interventions appear in this paper and must not be conflated. *(1) Artifact availability* — edit the realised message so z is present or absent, M_{-z} fixed. Estimand: marginal value of availability in an otherwise fixed handoff. Used in Experiment A. *(2) Policy* — instruct the relay to produce a message omitting z. A different estimand, since the relay may compensate elsewhere; reported only as sensitivity (Appendix C.4). *(3) Sham edit* — apply the editing machinery while preserving the value of z. Estimates the artefact itself. *(4) Fixed-skeleton prior* — hold H_{-z} and vary only the identity–value mapping. Estimand: prior access (Experiment B). *(5) Budget-neutral displacement* — insert z while evicting a matched-length atom. Estimand: Δ^{bu} (Experiment C).

A.3 EXPERIMENT-DEPENDENCY TABLE

Table 5: **Every definition and assumption is consumed by something.** Reading the table left to right traces a formal object forward to the experiment that uses it and the diagnostic that checks it; reading right to left traces an empirical design choice back to the result that requires it.

Formal object	Needed by	Observable diagnostic	Derivation
Def. 3.1	Exp. A, B, C	atom presence by canonical match	App. B.1
Asm. 3.2	Exp. A	non-focal diff; redundancy arm	App. B.3
Asm. 3.3	Exp. A, C	logged π_{iz}; stratum freq.	App. B.4
Asm. 3.5	Exp. A	sham placebo (mechanical only)	App. B.5
Asm. 3.4	Exp. A	pinned decoder; repeat count m	App. D.2
Prop. 4.1	Exp. A	terms sum to measured A	App. B.7
Thm. 4.4 (i)	motivation	— (impossibility)	App. B.8
Thm. 4.4 (ii)	Exp. A	one counterfactual query/atom	App. B.8
Cor. 4.5	Exp. A	R^{obs} vs $\bar{R}_0$	App. B.9
Cor. 4.6	Exp. B	$\bar{T}_{\mathrm{nat}} - \bar{T}_{\mathrm{blk}}$	App. B.10
Δ^{av} vs Δ^{bu}	Exp. C gate	displacement validation	App. B.13
Hop indexing	Exp. depth	$A^{(h)}$ vs $\bar{T}^{(h)}$	App. C.6

A.4 EXPANDED ALGORITHM 1: IMPLEMENTATION LOGIC

Atomisation and canonicalisation. Tier-1 values are normalised (case, whitespace, punctuation, number and date formats) before both inventory construction and matching, so presence detection is symmetric between source and message. **Edge cases in the editor.** If z appears more than once in M^{obs}, all occurrences are removed for M^{-z} and the multiplicity is logged. If z's natural surface form is entangled with a non-focal atom in one clause, the item is flagged and excluded, with the exclusion

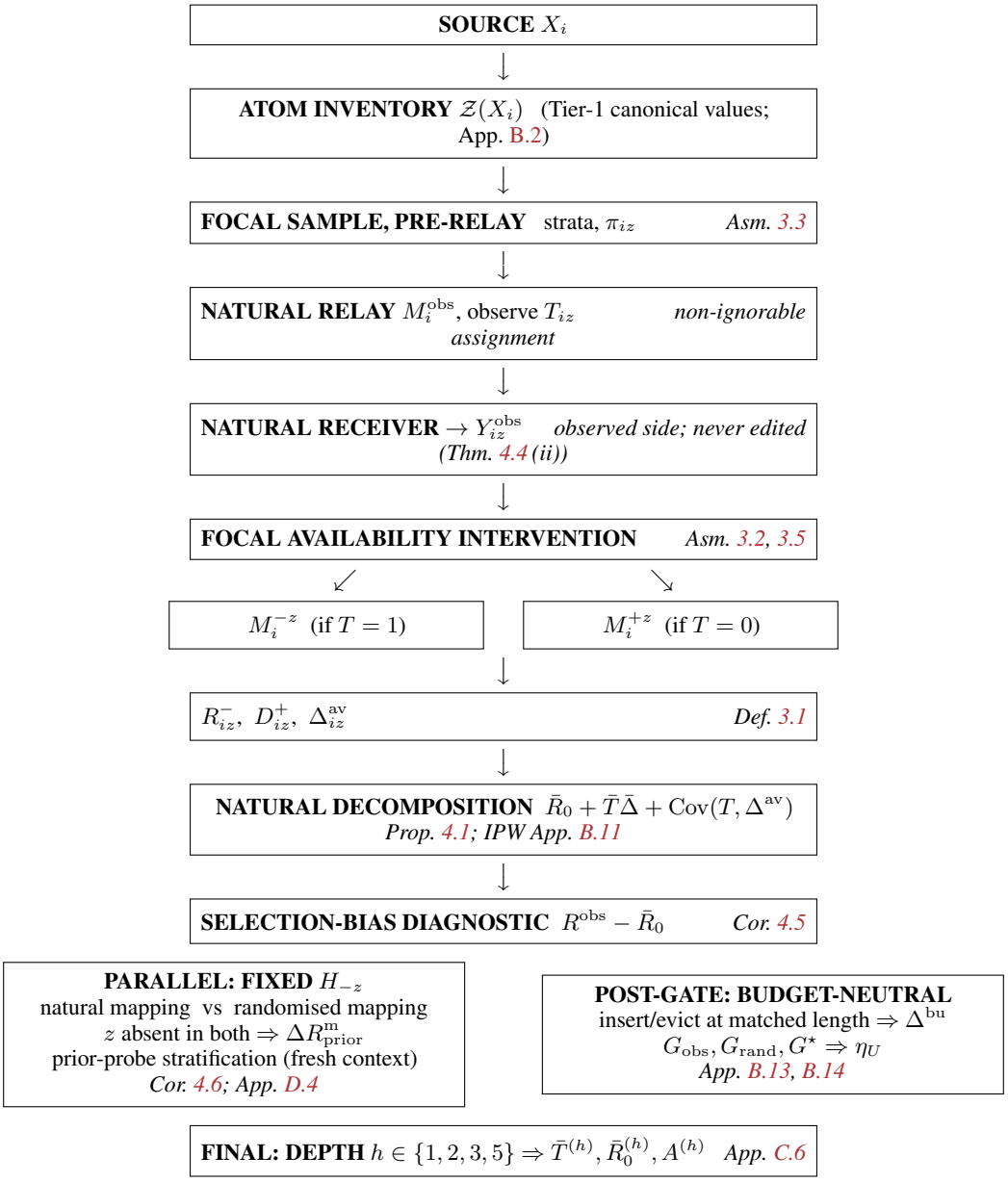

Figure 6: **End-to-end workflow with the licensing condition on every arrow.** The vertical spine is Experiment A; the left panel is Experiment B, which requires a fixed skeleton because Corollary 4.6 shows the natural twins do not identify prior access; the right panel is Experiment C, gated because $\Delta^{\text{av}} \neq \Delta^{\text{bu}}$. Nothing in the spine edits the observed side.

Table 6: **Mechanism-diverse comparator census.** The primary causal benchmark counts method families, not venues. Endpoint/hybrid and topology methods remain scientifically relevant but receive only metrics defined for their representation/action space.

Method	Publication	Representation/family	Role
LLMLINGUA-2	ACL Findings 2024	hard text; token classification	Primary causal-textual candidate
DAC	ACL 2025	hard text; attention–entropy token selection	Primary causal-textual candidate
LONGLLMLINGUA	ACL 2024	hard text; query-conditioned perplexity	Primary causal-textual candidate
Perception Compressor	NAACL Findings 2025	hard text; hierarchical query-aware	Primary causal-textual candidate
TACO-RL	ACL Findings 2025	hard text; task-reward token policy	Mandatory reproduce-or-explicitly-unavailable candidate
PROVENCE	ICLR 2025	text; cross-encoder context pruning	Candidate subject to usage/reproduction contract
RECOMP-EXT	ICLR 2024	hard text; bi-encoder/extractive	Historical textual anchor
IterCOMP	ACL 2026	generated hard text; iterative reasoning-aware	Adjacent multi-hop textual reference
COMI	ICLR 2026	latent merged memory	Endpoint/resource frontier
RAM	ACL 2026	hybrid text + vectors	Endpoint/resource frontier
GMSA	ACL 2026	soft tokens	Endpoint/resource frontier
SARA	ACL 2026	hybrid text + semantic vectors	Endpoint/resource frontier
AgentPrune	ICLR 2025	multi-agent communication topology	Adjacent multi-agent communication

rate reported. For M^{+z} the insertion point is the position the relay used for the same role in matched documents, defaulting to clause-final. **Budget handling.** Insertion overruns the budget by the inserted length; the overrun distribution is reported. Budget-neutral swaps occur only in Experiment C and never in the decomposition. **Placebo handling.** The sham subset is drawn independently of T so δ_{edit} is not itself selected on transmission. **Matched-skeleton construction.** $\mathcal{H}$ fixes the slot pattern of M_{-z} across arms: identical slots present, identical slots absent, matched positions and token lengths, differing only in the identity–value mapping. **Discovery-to-targeting chronology.** Under the original preregistration, Experiment C was licensed only after the discovery criterion. The completed discovery experiment did not meet that criterion, so no targeting result was produced under the original branch. A separate prospective protocol was frozen before any model-selection outcome and licenses Experiment C directly as a new budget-neutral targeting question. Its estimand remains Δ^{bu}; Δ^{av} is never substituted under a budget constraint.

A.5 BENCHMARK COMPARATOR ELIGIBILITY AND FAIRNESS

The benchmark is representation-stratified and mechanism-diverse. A *primary causal-textual* comparator must: (i) accept the same textual source and global downstream task; (ii) emit inspectable natural-language text in which canonical atom presence is defined; (iii) admit an externally enforced budget; (iv) use no focal-atom identity or intervention outcome; (v) have a peer-reviewed publication; and (vi) satisfy the prospectively frozen implementation, usage and native-task reproduction contract. Failure of a technical or usage criterion is reported; the row is not replaced after benchmark outcomes are observed.

Non-SOTA controls are load-bearing. Full context is a reference rather than a matched-budget competitor. Head truncation and Uniform-atoms are deterministic allocation floors; Uniform uses the exact CAUSALRELAY renderer. A vanilla relay is the natural generative baseline. The source-salience selector uses no outcome label. The role-only utility relay measures how much is explained by atom class. Natural-policy imitation predicts T from the same source-only feature surface, isolating behaviour cloning from causal-value learning. When the already licensed availability labels suffice, an availability-value relay predicts Δ^{av} and uses the same knapsack/renderer, providing a direct control for the distinction $\Delta^{\text{av}} \neq \Delta^{\text{bu}}$.

Common task and budget interface. Query-aware methods receive the same global downstream task, never sampled focal identities or values. Every hard-text method must satisfy $L(M_j) \leq B$ under a common tokenizer; exact equality is not forced when a method's selection unit makes it impossible without destructive post-processing. Budget-controlling hyperparameters are selected on the model-selection split before evaluation and frozen. Realised token counts are always reported.

Reproduction and usage contract. Before our benchmark is revealed, each candidate records paper identifier, official repository, code revision, model/checkpoint revision, licence/usage classification, environment lock, native benchmark and expected metric. It must reproduce a prespecified native/public result within the frozen tolerance and pass output inspectability, task-information, budget and no-leakage tests. Trainable comparators may use only the same model-selection information licensed to them by their original method; no comparator is trained on held-out outcomes.

Family-aware simultaneous superiority. Family labels are fixed from mechanism/backbone/training lineage before our benchmark. If a family has multiple eligible members, a family win requires CAUSALRELAY to qualify against all of them. At least four distinct families must be ready for any broad causal-textual superiority headline. With exactly four, all four must be family wins; with five or more, at least four distinct families must be wins. For every comparator j, both simultaneous one-sided lower bounds must be positive:

$$L_{0.95}^{\mathrm{sim}}(A_{\mathrm{ours}} - A_j) > 0, \qquad L_{0.95}^{\mathrm{sim}}(C_{\mathrm{comm,ours}} - C_{\mathrm{comm},j}) > 0.$$

The bounds come from a max-statistic paired document-cluster bootstrap over the full primary comparator–metric family with 10,000 replicates. This is stricter than counting several correlated pointwise 95% intervals. Failure yields the complete table without superiority wording.

Endpoint-frontier veto. A latent/hybrid method cannot enter T-based inference, but it can falsify a claim of generic context-compression state of the art. If an eligible contemporary endpoint method Pareto-dominates CAUSALRELAY on endpoint fidelity and resource use at matched effective context, any positive headline is scoped to inspectable causal-textual handoff allocation. No representation class is silently dropped because it performs well.

B PROOFS AND THEORETICAL DERIVATIONS

This appendix separates mathematical identities, identification results, assumption-dependent consequences and estimator derivations.

B.1 FORMAL PROBABILITY SPACE AND INDEXING

Let $(\Omega, \mathcal{F}, \mathbb{P})$ carry four independent sources of randomness: (i) document draw $i \sim \mathcal{D}$ over a superpopulation of sources; (ii) a focal-*selection* vector $S_i = (S_{iz})_{z \in \mathcal{Z}(X_i)}$ from a fixed-size role-stratified design without replacement ($\sum_z S_{iz} = k$) with first-order inclusion probabilities $\pi_{iz} > 0$, drawn independently of (iii) and (iv); (iii) relay randomness U^S; (iv) receiver randomness U^R, drawn afresh for every query. Potential outcomes $Y_{iz}(w; M_{i,-z})$ are $\{0,1\}$-valued functions of $(w, M_{i,-z}, U^R)$, and R_{iz}^-, D_{iz}^+ of Definition 3.1 are conditional expectations over U^R with $(i, z, M_{i,-z}^{\mathrm{obs}})$ held fixed — so they are random variables with respect to (i)–(iii) and deterministic given the realised message. Under decoder regime (D) of Assumption 3.4 the conditioning collapses: $Y_{iz}(w; \cdot)$ is a fixed $\{0,1\}$-valued function and $R_{iz}^-, D_{iz}^+ \in \{0,1\}$, which is what makes atom-level identification possible from a single query (Theorem 4.4 (ii-a)). Under regime (S) they are genuine expectations and a single query is an unbiased draw, not an identification of the mean.

It is important that nothing here asserts $z \perp T$. The relay transmits selectively, so atom identity and transmission are dependent by construction; what Assumption 3.3 supplies is that the *selection* of atoms into our sample is independent of that dependence, which is precisely what a design-based (Horvitz–Thompson) argument requires.

Three levels of aggregation are used and never conflated. *Atom level*: $R_{iz}^-, D_{iz}^+, \Delta_{iz}^{\mathrm{av}}, T_{iz}$. *Document level*: weighted means over $z \in \mathcal{Z}(X_i)$ with weights $1/\pi_{iz}$. *Population level*: expectations over $\mathcal{D}$. All displayed estimands $(\bar{R}_0, \bar{T}, \bar{\Delta}, A)$ are population-level under the design measure $\mathbb{P} \otimes \Pi$. Because outcomes lie in $[0,1]$, all first and second moments exist and $\mathrm{Cov}(T, \Delta^{\mathrm{av}})$ is finite, so Eq. (3) requires no integrability side condition. Positivity $\pi_{iz} > 0$ is required for the Horvitz–Thompson form of Appendix B.11 and is guaranteed by construction because strata are defined on the inventory before sampling.

A finite-population reading is also available: fix the realised set of documents and atoms and take all expectations over (ii)–(iv) only. Every result below holds verbatim under either reading; the

superpopulation reading is used when generalising beyond the sampled corpus, and we flag the distinction wherever the variance formula differs (Appendix D.1).

B.2 Why the epistemic atom is the unit of intervention

An availability intervention must be applied to something that can be present or absent from a message without altering anything else. *Tier-1* atoms — entities, scope labels, periods, numerics, provenance references — have canonical values, so presence is decidable by normalised exact match and the edit is localised. *Tier-2* atoms — relation strength, certainty, counterevidence — are paraphrasable, so both presence detection and minimal editing become model-mediated, which would put a judge inside the primary outcome. The causal core is therefore Tier-1 only, and Tier-2 is a post-discovery extension. The limitation is real and stated in the main text: our estimands concern the transmission of canonically valued content, not of hedging.

B.3 Assumption 3.2: consistency and confined interference

Formal statement. For every (i, z) and $w \in \{0, 1\}$, $Y_{iz}^{\text{obs}} = Y_{iz}(T_{iz}; M_{i,-z}^{\text{obs}})$, and the map $\mathcal{E}$ realising $W_{iz} = w$ satisfies $\mathcal{E}(M_i, z, w)_{-z} = M_{i,-z}$. **Why needed.** It supplies the pointwise identity underlying Proposition 4.1; without it Y^{obs} is not a potential outcome at the realised treatment. **Consumed by.** Proposition 4.1, Theorem 4.4. **Observable.** M_i and its atom content are directly inspectable, so realised availability is measured, not inferred; the editor's non-focal invariance is checkable by diffing $M_{i,-z}^{\pm z}$ against $M_{i,-z}^{\text{obs}}$. **Not observable.** Whether the receiver's internal use of M_{-z} changes when z is removed. **Failure mode.** Redundancy: a non-focal atom carries information about z, so $Y_{iz}(0)$ depends on $M_{i,-z}$. This does not break the estimand — $M_{i,-z}$ is held at its observed value, so R^- remains well defined — but it makes R^- a *conditional* quantity, not an intrinsic reconstructability. **Diagnostic and sensitivity.** The redundancy-ablated arm (Table 14) removes within-message corroborating cues and reports the shift in R^-; a large shift means in-context redundancy, not parametric prior, is the operative recovery route.

B.4 Assumption 3.3: focal-atom exogeneity and positivity

Formal statement. S_{iz} is the indicator of a fixed-size role-stratified sample without replacement of exactly k atoms from $\mathcal{Z}(X_i)$, with first-order inclusion probability $\pi_{iz} = \mathbb{P}[S_{iz} = 1 \mid \cdot] > 0$ measurable with respect to $\sigma(X_i, \text{role}(z))$ and drawn independently of U^S, U^R; hence Eq. (2). **What is *not* assumed.** We do not assume $z \perp T$, which is false: a relay transmits atoms according to their properties, so identity and transmission are dependent, and that dependence is exactly what $\text{Cov}(T, \Delta^{\text{av}})$ measures. The independence we need is of the *selection indicator*, which makes the focal sample a probability sample of the inventory and licenses the design-based estimators of Appendix B.11. Conflating the two would be a genuine error and an earlier draft of this manuscript contained it. **Why sampling from M is invalid.** Drawing focal atoms from the message makes S_{iz} a function of T_{iz}, violating Eq. (2), and conditions on $T_{iz} = 1$. The stratum $\{T = 0\}$ is then never sampled, $\text{Cov}(T, \Delta^{\text{av}})$ is undefined, and Proposition 4.1 collapses. The ordering *inventory* $\rightarrow$ *sample* $\rightarrow$ *relay* is therefore not a convenience but a requirement, and it is the single most easily violated step in an implementation. **Role-stratified sampling.** Balancing draws across roles is permitted and desirable for power, but then π_{iz} differs by stratum and unweighted means estimate a stratum-balanced population rather than the atom population. Report stratum-specific estimates, or apply the weights of Appendix B.11; do not describe a balanced diagnostic sample as an unbiased sample of the inventory. **Diagnostic.** Log π_{iz} per draw and verify the realised stratum frequencies against the design.

B.5 Assumption 3.5: edit neutrality

Formal statement. Writing $Y^{\mathcal{E}}$ for the outcome on an edited artefact, $Y_{iz}^{\mathcal{E}}(w) = Y_{iz}(w)$ for $w = 1 - T_{iz}$. **Why deletion and insertion are asymmetric.** For $T = 1$ the counterfactual is produced by *deletion*, which can leave a syntactic gap; for $T = 0$ it is produced by *insertion*, which must place new content and may perturb length. These are different operations and there is no reason for their artefacts to cancel. **Why we do not construct both sides.** Editing the observed side as well would make the estimand Δ^{av} on *edited* messages, so Proposition 4.1 would decompose the

endpoint fidelity of an edited pipeline rather than the deployed one (Appendix B.8). We therefore keep the natural artefact and test the assumption instead. **What the placebo does and does not do.** The sham edit applies the editing machinery while preserving the semantic content of z, so it isolates the *mechanical* component of the artefact: formatting disruption, fluency loss, length change, position shift. It is a negative control on that component and nothing more. It cannot establish the exclusion restriction, because the quantity at issue — whether removing the *meaning* of z affects the receiver only through the availability of z — is by construction not varied by a semantics-preserving edit. Claiming the placebo "tests Assumption 3.5" would be an overstatement, and an earlier draft made it. **Where the burden actually sits.** The partial-identification analysis of Appendix D.3 is primary: it reports a region for Δ^{av} under an assumed bound on the residual artefact, with the point correction recovered only under an additional additive model that we state rather than presume. A coherence audit reports the exclusion rate for edited artefacts.

B.6 Mathematical conventions

All proofs use the probability space of Appendix B.1. Outcomes are $\{0,1\}$-valued, so every expectation below is of a bounded random variable and all first and second moments exist; no integrability side conditions are needed. Unless stated otherwise, $\mathbb{E}$ denotes expectation under the design measure $\mathbb{P} \otimes \Pi$ jointly over document draw, focal-atom draw, relay randomness and receiver randomness. Conditional expectations $R^{-}_{iz} = \mathbb{E}[Y_{iz}(0) \mid i, z, M^{\mathrm{obs}}_{i,-z}]$ average over receiver randomness only. We write $\bar{T} = \mathbb{E}[T]$ and assume $\bar{T} \in (0,1)$ wherever a conditional on $\{T=0\}$ or $\{T=1\}$ appears; this is a genuine positivity requirement and is stated in each result that uses it. Under role-stratified sampling, unweighted sample means estimate the stratum-balanced population; the Horvitz–Thompson corrections are Appendix B.11 and are suppressed in the proofs for readability.

B.7 Proof of Proposition 4.1

Proof. Step 1 (pointwise identity). Fix (i,z) and condition on $M^{\mathrm{obs}}_{i,-z}$. By Assumption 3.2 the realised outcome equals the potential outcome at the realised availability, $Y^{\mathrm{obs}}_{iz} = Y_{iz}(T_{iz}; M^{\mathrm{obs}}_{i,-z})$. Since $T_{iz} \in \{0,1\}$, the algebraic identity

$$Y_{iz}(T_{iz}) = Y_{iz}(0) + T_{iz}\big[Y_{iz}(1) - Y_{iz}(0)\big]$$

holds pointwise in U^R. Taking $\mathbb{E}[\,\cdot \mid i, z, M^{\mathrm{obs}}_{i,-z}]$ and using Definition 3.1 gives $\mathbb{E}[Y^{\mathrm{obs}}_{iz} \mid \cdot] = R^{-}_{iz} + T_{iz}\Delta^{\mathrm{av}}_{iz}$, where T_{iz} is measurable with respect to the conditioning σ-field.

Step 2 (expectation under the design). Let $\mathcal{N}$ be the target atom population and, for an atom-level quantity g_{iz}, write $\mathbb{E}[g] = |\mathcal{N}|^{-1}\sum_{(i,z)\in\mathcal{N}} g_{iz}$ for its population mean. By Assumption 3.3 the selection indicators satisfy Eq. (2) with $\pi_{iz} > 0$, so $\mathbb{E}\big[|\mathcal{N}|^{-1}\sum_{(i,z)} \pi^{-1}_{iz} S_{iz} g_{iz}\big] = \mathbb{E}[g]$ for any g measurable with respect to $(X_i, \mathcal{Z}(X_i), U^S, U^R)$ — in particular for $g \in \{R^{-}, \Delta^{\mathrm{av}}, T, T\Delta^{\mathrm{av}}\}$. Applying this to the Step-1 identity gives $A = \mathbb{E}[R^{-}] + \mathbb{E}[T\Delta^{\mathrm{av}}] = \bar{R}_0 + \mathbb{E}[T\Delta^{\mathrm{av}}]$, where every expectation is over the atom population and *not* over the relay's transmission decisions. Note that T and Δ^{av} are dependent — that is the content of the third term — and no independence between z and T is used or available.

Step 3 (covariance expansion). Both factors are bounded, so $\mathbb{E}[T\Delta^{\mathrm{av}}] = \mathbb{E}[T]\mathbb{E}[\Delta^{\mathrm{av}}] + \mathrm{Cov}(T, \Delta^{\mathrm{av}}) = \bar{T}\bar{\Delta} + \mathrm{Cov}(T, \Delta^{\mathrm{av}})$, which gives Eq. (3). No approximation is used at any step, so the identity is exact; $C_{\mathrm{comm}} = A - \bar{R}_0$ follows by rearrangement.

Step 4 (stratified form). If atoms are selected with unequal π_{iz}, replace each population mean by a design-weighted analogue. The *Horvitz–Thompson* estimator is $\widehat{\mathbb{E}}_{\mathrm{HT}}[g] = |\mathcal{N}|^{-1}\sum_{(i,z)\in\mathcal{S}} \pi^{-1}_{iz} g_{iz}$, which is unbiased but requires the population size $|\mathcal{N}|$; the *Hájek* estimator is the self-normalised ratio $\widehat{\mathbb{E}}[g] = \big(\sum_{\mathcal{S}} \pi^{-1}_{iz} g_{iz}\big)/\big(\sum_{\mathcal{S}} \pi^{-1}_{iz}\big)$, which is ratio-consistent rather than exactly unbiased but does not require $|\mathcal{N}|$. These are distinct estimators and we use the Hájek form throughout, since $|\mathcal{N}|$ is known only up to atomiser behaviour. Here $\mathcal{S}$ is the realised focal sample. The identity holds within each stratum without weighting, and across strata after weighting; positivity $\pi_{iz} > 0$ is what makes the weights finite. Details: Appendix B.11.

Step 5 (sign). Nothing above requires $\Delta_{iz}^{\mathrm{av}} \geq 0$. If $\Delta_{iz}^{\mathrm{av}} < 0$ for some atoms, C_{comm} may be negative and Eq. (3) is unaffected; only the interpretation of $\phi = C_{\mathrm{comm}}/A$ as a proportion fails, which is why it is reported as a signed quantity (Remark 4.3). □

Honest appraisal. Steps 1–3 are the law of total expectation for a binary treatment. We label the result a Proposition for this reason. Its value is not algebraic: it is that Theorem 4.4 makes each of the three terms separately identifiable, and that each is then reported as a distinct empirical quantity in Table 1.

B.8 Proof of Theorem 4.4

Proof. (i) Non-identification. Because Y^{obs} and T are both binary, the observed law $\mathcal{O}_{\mathrm{end}}$ is fully determined by the three numbers

$$\bar{T} = \mathbb{P}(T = 1), \qquad d_1 = \mathbb{P}(Y^{\mathrm{obs}} = 1 \mid T = 1), \qquad r_0 = \mathbb{P}(Y^{\mathrm{obs}} = 1 \mid T = 0).$$

By Assumption 3.2, $d_1 = \mathbb{E}[D^+ \mid T = 1]$ and $r_0 = \mathbb{E}[R^- \mid T = 0]$. The two remaining conditional means,

$$r_1 := \mathbb{E}[R^- \mid T = 1], \qquad d_0 := \mathbb{E}[D^+ \mid T = 0],$$

are never realised in an uninstrumented run: for a transmitted atom the run does not reveal what would have happened had it been omitted, and symmetrically for an omitted atom. In a nonparametric model no constraint links them to $(\bar{T}, d_1, r_0)$, so (r_1, d_0) may take any value in $[0,1]^2$. The realising construction must be valid in both decoder regimes, and in particular under regime (D) where atom-level $R_{iz}^-, D_{iz}^+ \in \{0,1\}$, so a degenerate atom-level distribution at a non-integer mean is *not* available. Use instead a Bernoulli mixture across the atom superpopulation: within the stratum $\{T = 1\}$ let a sub-population of mass r_1 have $R_{iz}^- = 1$ and the complement have $R_{iz}^- = 0$, so the conditional mean is exactly r_1 while every atom-level value remains binary; construct d_0 on $\{T = 0\}$ identically. For a finite population of N_1 atoms in the stratum this realises any r_1 on the grid $\{0, 1/N_1, \ldots, 1\}$, and the full interval $[0,1]$ is recovered either in the superpopulation limit or by closure, since the induced map from mixture weights to $(\bar{R}_0, \bar{\Delta}, \mathrm{Cov})$ is continuous. Under regime (S) the same construction works a fortiori, atom-level values then being free in $[0,1]$.

Expressing the three terms of Eq. (3) in these five numbers,

$$\begin{aligned} \bar{R}_0 &= \bar{T}\, r_1 + (1 - \bar{T})\, r_0, \\ \bar{\Delta} &= \bar{T}\,(d_1 - r_1) + (1 - \bar{T})\,(d_0 - r_0), \\ \mathrm{Cov}(T, \Delta^{\mathrm{av}}) &= \bar{T}(1 - \bar{T})\big[(d_1 - r_1) - (d_0 - r_0)\big]. \end{aligned} \tag{6}$$

For $\bar{T} \in (0,1)$ the partial derivatives are $\partial \bar{R}_0/\partial r_1 = \bar{T} \neq 0$, $\partial \bar{\Delta}/\partial d_0 = 1 - \bar{T} \neq 0$ and $\partial \mathrm{Cov}/\partial r_1 = -\bar{T}(1 - \bar{T}) \neq 0$, so all three terms are non-constant affine functions of (r_1, d_0). Hence the two-parameter family indexed by $(r_1, d_0) \in [0,1]^2$ consists of causal worlds inducing *exactly the same* $\mathcal{O}_{\mathrm{end}}$ — not merely the same A — while realising distinct $\bar{R}_0$, $\bar{\Delta}$ and $\mathrm{Cov}(T, \Delta^{\mathrm{av}})$.

As an internal check, substituting Eq. (6) into Eq. (3) gives

$$\bar{R}_0 + \bar{T}\bar{\Delta} + \mathrm{Cov}(T, \Delta^{\mathrm{av}}) = \bar{T} r_1 + (1 - \bar{T}) r_0 + \bar{T}(d_1 - r_1) = (1 - \bar{T}) r_0 + \bar{T} d_1 = A,$$

identically in (r_1, d_0). The decomposition therefore holds in every member of the family, which is precisely why A carries no information about the split.

(ii) Identification. Select z per Assumption 3.3 and run the relay, yielding T_{iz} and M_i^{obs}. Querying the receiver on the natural M_i^{obs} realises, by Assumption 3.2, the coordinate indexed by T_{iz}. Constructing the single counterfactual artefact with $W_{iz} = 1 - T_{iz}$ and $M_{i,-z} = M_{i,-z}^{\mathrm{obs}}$ and querying again realises the complementary coordinate, by Assumption 3.5. What this yields depends on the decoder regime of Assumption 3.4.

(ii-a) Regime (D). $Y_{iz}(w; \cdot)$ is a deterministic $\{0,1\}$-valued function, so the two queries *are* R_{iz}^- and D_{iz}^+; both coordinates are observed exactly and $\Delta_{iz}^{\mathrm{av}} = D_{iz}^+ - R_{iz}^-$ is identified at atom level. Design-weighted averaging then identifies $\bar{R}_0$ and $\bar{\Delta}$, pairing with T_{iz} identifies $\mathrm{Cov}(T, \Delta^{\mathrm{av}})$, and Proposition 4.1 reproduces the measured A as an internal consistency check.

(ii-b) Regime (S). Write the realised queries as $\tilde{R}^-_{iz} = R^-_{iz} + \xi^0_{iz}$ and $\tilde{D}^+_{iz} = D^+_{iz} + \xi^1_{iz}$ with $\mathbb{E}[\xi^w_{iz} \mid i, z, M^{\text{obs}}_{i,-z}] = 0$ and ξ independent across queries. A single Bernoulli draw cannot identify its own mean, so R^-_{iz}, D^+_{iz} and Δ^{av}_{iz} are *not* identified at atom level; m independent draws per availability identify them only as $m \to \infty$, at rate $O_p(m^{-1/2})$. The population estimands survive, however. Writing $\tilde{\Delta^{\text{av}}}_{iz} = \Delta^{\text{av}}_{iz} + \xi^1_{iz} - \xi^0_{iz}$, design-weighted averaging gives $\mathbb{E}[\tilde{\Delta^{\text{av}}}] = \bar{\Delta}$ and $\mathbb{E}[\tilde{R}^-] = \bar{R}_0$. For the third term, T_{iz} is measurable with respect to $(X_i, \mathcal{Z}(X_i), U^S)$ and the receiver noise has conditional mean zero given that σ-field, so

$$\operatorname{Cov}\big(T, \tilde{\Delta^{\text{av}}}\big) = \operatorname{Cov}(T, \Delta^{\text{av}}) + \mathbb{E}\big[T\,\mathbb{E}[\xi^1 - \xi^0 \mid \cdot]\big] - \bar{T} \cdot 0 = \operatorname{Cov}(T, \Delta^{\text{av}}).$$

Hence the estimating equations built from one query per availability are unbiased for $\bar{R}_0$, $\bar{\Delta}$ and $\operatorname{Cov}(T, \Delta^{\text{av}})$ even under a stochastic receiver, so these population functionals are identified *at the level of the distribution*. We are careful not to overstate this: a single noisy draw per atom does not numerically recover a population expectation in finite samples. Consistency requires the number of sampled atoms and documents to grow, the finite-sample estimator retains the receiver-noise variance $\operatorname{Var}(\xi^1 - \xi^0)$ inflating its sampling variance by $O(1/n)$ relative to regime (D), and only the atom-level quantities require $m \to \infty$ repeats. This is why the primary protocol pins the decoder (Appendix D.10) and why atom-level quantities such as $\mathbb{P}[\Delta^{\text{av}}_{iz} < 0]$ are reported only under regime (D) or with explicit repeat counts.

The observed side must remain the natural artefact. Were both sides constructed, the identified quantity would be Δ^{av} defined on edited messages, and Proposition 4.1 would decompose the endpoint fidelity of an edited pipeline rather than of the deployed one. Assumption 3.5 is then needed only for the single constructed side, and its mechanical component is bounded by the placebo while its semantic component is carried through the partial-identification region (Appendix D.3).

(iii) Atom-level minimality under regime (D). Work in a nonparametric query model: a design specifies, for each sampled atom, a set of receiver queries at chosen availabilities, and places no restriction linking distinct atoms. Under regime (D) each atom carries the pair $\big(Y_{iz}(0), Y_{iz}(1)\big) \in \{0,1\}^2$ and the natural run reveals exactly the coordinate indexed by T_{iz}. Fix an atom (i, z) and suppose the design issues no counterfactual query *at that atom*. By the construction in part (i), applied at (i, z) with all other atoms held fixed, there are two causal worlds agreeing on every quantity the design observes and differing in the unobserved coordinate, hence in Δ^{av}_{iz}. So Δ^{av}_{iz} is not identified, and one counterfactual query at (i, z) reveals the missing coordinate and suffices by part (ii-a). This is the exact sense in which the per-atom query is minimal, and it is an *atom-level* statement.

What does not follow. It does not follow that one query per sampled atom is minimal for the population quantities. Zero counterfactual information anywhere leaves $(\bar{R}_0, \bar{\Delta}, \operatorname{Cov}(T, \Delta^{\text{av}}))$ unidentified — the family of part (i) moves all three while fixing $\mathcal{O}_{\text{end}}$ — but this is an argument against *zero* intervention, not for *universal* intervention. Let Q_{iz} be a second-phase intervention indicator drawn independently with known $\mathbb{P}[Q_{iz} = 1 \mid \cdot] = q_{iz} > 0$, and query counterfactually only when $Q_{iz} = 1$. Two-phase design weighting then identifies and consistently estimates the population decomposition without instrumenting every focal atom (Appendix B.12). An earlier draft of this manuscript asserted population-level minimality; that claim was too strong and is withdrawn.

We claim minimality only in this model. Under additional structure — for instance a parametric or smoothness assumption linking atoms, or a model in which Δ^{av} depends only on observed atom covariates — fewer than one query per atom could suffice, and we make no claim to the contrary. □

B.9 Proof of Corollary 4.5

Proof. Assume $\bar{T} \in (0, 1)$ so that both conditionals are well defined. Decomposing the marginal mean,

$$\bar{R}_0 = \mathbb{E}[R^-] = \bar{T}\,\mathbb{E}[R^- \mid T = 1] + (1 - \bar{T})\,\mathbb{E}[R^- \mid T = 0],$$

and subtracting from $R^{\text{obs}} = \mathbb{E}[R^- \mid T = 0]$,

$$R^{\text{obs}} - \bar{R}_0 = \bar{T}\big(\mathbb{E}[R^- \mid T = 0] - \mathbb{E}[R^- \mid T = 1]\big).$$

For a binary T and any bounded V, $\operatorname{Cov}(T, V) = \bar{T}(1 - \bar{T})\big(\mathbb{E}[V \mid T = 1] - \mathbb{E}[V \mid T = 0]\big)$. Applying this with $V = R^-$ and substituting gives Eq. (4). Since $1 - \bar{T} > 0$, the bias vanishes if and

only if $\mathrm{Cov}(T, R^-) = 0$; the "only if" direction requires exactly the positivity assumed, and fails to be meaningful at $T \in \{0, 1\}$ where one conditional is undefined.

Sign. A relay that preferentially omits reconstructable atoms induces $\mathrm{Cov}(T, R^-) < 0$ and hence $R^{\mathrm{obs}} > \bar{R}_0$: the observational estimator *overstates* reconstructability, and by an amount increasing in the relay's allocation skill.

Stratified form. Under unequal π_{iz} each conditional mean is replaced by its Hájek analogue within stratum and the identity holds stratum-wise; the pooled statement requires the weights of Appendix B.11, and an unweighted pooled estimate of $R^{\mathrm{obs}} - \bar{R}_0$ is biased whenever transmission rates differ across strata. □

B.10 Proof of Corollary 4.6

Proof. Let ρ denote the natural identity–value mapping assigning surface values to atom slots, and ρ' the randomised mapping produced by $\mathcal{T}$, which preserves slot structure, role, position and token length while replacing values. Write $M_{-z,\mathrm{nat}}$ for the non-focal content of the natural relay run and $M_{-z,\mathrm{blk}}$ for that of the randomised twin, and let $M'_{-z,\mathrm{nat}}$ denote the natural transmission pattern carrying the randomised mapping — that is, ρ' applied to the slots that $M_{-z,\mathrm{nat}}$ actually contains.

Write $\mu_z(w; M) = \mathbb{E}_{U^R}[Y_z(w; M)]$ for the receiver-mean response, which under regime (D) coincides with $Y_z(w; M)$ itself and under regime (S) is the conditional expectation of Definition 3.1; all displays below are in μ, so the identity is valid in both regimes. Adding and subtracting $\mu_z(0; M'_{-z,\mathrm{nat}})$,

$$R^-_{z,\mathrm{nat}} - R^-_{z,\mathrm{blk}} = \underbrace{\big[\mu_z(0; M_{-z,\mathrm{nat}}) - \mu_z(0; M'_{-z,\mathrm{nat}})\big]}_{\text{prior access}} + \underbrace{\big[\mu_z(0; M'_{-z,\mathrm{nat}}) - \mu_z(0; M_{-z,\mathrm{blk}})\big]}_{\text{context composition}}. \tag{7}$$

The first bracket varies only the mapping at fixed non-focal transmission pattern; the second varies the transmission pattern at fixed mapping. The second is non-zero in general: randomisation changes atom surprisal, so the relay's selection may differ between arms, i.e. $T_{-z,\mathrm{nat}} \neq T_{-z,\mathrm{blk}}$ with positive probability. Forcing $W_z = 0$ in both arms constrains only the focal atom and leaves the second bracket untouched, so the two terms remain non-separable from the natural twins alone.

Now suppose both arms are generated from a common skeleton $H_{-z} = \mathcal{H}(\cdot)$, so that by construction $M_{-z,\mathrm{nat}} = M_{-z,\mathrm{blk}} = H_{-z}$ up to the mapping. Then $M'_{-z,\mathrm{nat}} = M_{-z,\mathrm{blk}}$, the second bracket vanishes identically, and the observed contrast equals the first bracket. Hence

$$\Delta R^{\mathrm{m}}_{\mathrm{prior}} = \mathbb{E}\big[\mu_z(0; H_{-z}, \rho) - \mu_z(0; H_{-z}, \rho')\big]$$

is identified.

Scope of the claim. The argument establishes two things: that the naïve natural-versus-randomised contrast is *contaminated* by the context-composition term, and that a common skeleton is *sufficient* to remove it. It does not establish that a matched skeleton is the unique route to identification. Within the restricted design class that varies only the identity–value mapping and does not model the relay's selection policy, the second bracket cannot be removed by any other means, so necessity holds there; outside that class — for instance if the selection policy were modelled and the composition term adjusted for parametrically — other identification strategies may exist and we make no claim about them.

What is and is not identified. Experiment B identifies the effect of the identity–value mapping on focal-omission reconstruction *at the fixed skeleton* H_{-z}. It does not identify the prior-access effect under the relay's own natural transmission pattern, which is a different (and policy-dependent) estimand, and it does not identify Δ^{av}. The natural-versus-randomised difference in transmission rate, $\bar{T}_{\mathrm{nat}} - \bar{T}_{\mathrm{blk}}$, is therefore reported as a manipulation check on the second bracket rather than as part of the identification argument. □

This appendix derives the estimators and structural facts that determine how the experiments of Sections 5–6 are built. Each subsection closes with the design choice it forces.

B.11 Stratified focal sampling and weighted estimators

Let $\mathcal{S}$ be the realised focal sample with inclusion probabilities $\pi_{iz} > 0$. For an atom-level quantity g_{iz} the Horvitz–Thompson estimator is $\widehat{\mathbb{E}}_{\mathrm{HT}}[g] = |\mathcal{N}|^{-1} \sum_{(i,z)\in\mathcal{S}} \pi_{iz}^{-1} g_{iz}$ with $\mathcal{N}$ the target atom population, and the Hájek (self-normalised) form is

$$\widehat{\mathbb{E}}[g] = \frac{\sum_{(i,z)\in\mathcal{S}} \pi_{iz}^{-1} g_{iz}}{\sum_{(i,z)\in\mathcal{S}} \pi_{iz}^{-1}}. \tag{8}$$

We use the Hájek form throughout, since $|\mathcal{N}|$ is known only up to atomiser behaviour and the ratio is invariant to that. Applying Eq. (8) to $g \in \{R^-, \Delta^{\mathrm{av}}, T, T\Delta^{\mathrm{av}}\}$ and forming $\widehat{\mathrm{Cov}}(T, \Delta^{\mathrm{av}}) = \widehat{\mathbb{E}}[T\Delta^{\mathrm{av}}] - \widehat{\mathbb{E}}[T]\widehat{\mathbb{E}}[\Delta^{\mathrm{av}}]$ gives a weighted version of Eq. (3) that is consistent for the inventory population. If role strata are sampled at equal sizes while their inventory frequencies differ, unweighted means instead estimate the stratum-balanced population; the discrepancy is $\mathbb{E}[g] - \mathbb{E}_{\mathrm{bal}}[g] = \sum_s (w_s - b_s)\mathbb{E}[g \mid s]$ over strata s with inventory shares w_s and design shares b_s, which is non-zero exactly when g varies across roles — and R^- does, strongly. **Experimental consequence:** report both stratum-specific estimates and the weighted pooled estimate (Table 7, Table 8), and never a single unweighted pooled number.

B.12 Two-phase intervention sampling

Theorem 4.4 (iii) scopes minimality to atom-level effects, because the population decomposition does not require a counterfactual query at every sampled atom. Let $Q_{iz} \in \{0, 1\}$ be a second-phase intervention indicator drawn independently of everything else with known $\mathbb{P}[Q_{iz} = 1 \mid \mathcal{F}] = q_{iz} > 0$, and issue the counterfactual query only when $Q_{iz} = 1$. For an atom-level quantity g_{iz} observed only on the intervened subsample, the two-phase Horvitz–Thompson estimator is

$$\widehat{\mathbb{E}}_{\mathrm{HT}}[g] = \frac{1}{|\mathcal{N}|} \sum_{(i,z)} \frac{S_{iz} Q_{iz}}{\pi_{iz}\, q_{iz}}\, g_{iz}, \tag{9}$$

with the Hájek analogue obtained by self-normalising by $\sum S_{iz}Q_{iz}/(\pi_{iz}q_{iz})$. Since $Q \perp (T, R^-, D^+)$ by construction and $q_{iz} > 0$ (positivity at the second phase), Eq. (9) is unbiased for $\mathbb{E}[g]$; applying it to $g \in \{R^-, \Delta^{\mathrm{av}}, T\Delta^{\mathrm{av}}\}$ identifies and consistently estimates all three terms of Eq. (3). The cost is variance: the second-phase inflation is of order $\mathbb{E}[(1-q)/q]$ relative to full instrumentation, which is why the discovery sample sets $q_{iz} \equiv 1$. **Experimental consequence:** full instrumentation is a precision choice; the minimality claim in the main text is stated only at atom level, and any future cost-constrained replication may use $q_{iz} < 1$ without losing population identification.

B.13 Availability versus fixed-budget effects

Define, for a budget B and a relay allocation $\mathbf{t}$ with $\sum_z \ell_z t_z \le B$,

$$\Delta_z^{\mathrm{av}} = Y_z(1; M_{-z}) - Y_z(0; M_{-z}), \qquad \Delta_z^{\mathrm{bu}} = Y_z(1; M_{-z\setminus e}) - Y_z(0; M_{-z}),$$

where e is a matched-length atom evicted to make room. The two differ by the eviction's own effect: $\Delta_z^{\mathrm{av}} - \Delta_z^{\mathrm{bu}} = Y_z(1; M_{-z}) - Y_z(1; M_{-z\setminus e})$, which is zero only if removing e does not affect recovery of z. That holds when e carries no information about z and does not alter the receiver's use of the remaining context — a condition that fails precisely under the redundancy identified in Appendix B.3. Consequently a fixed-budget oracle fed Δ^{av} optimises the wrong objective, and any targeting claim built on Δ^{av} is open to the objection that its effects were estimated under a violated constraint. **Experimental consequence:** Experiment C uses Δ^{bu} exclusively, is gated, and validates that the Δ^{av} ranking survives displacement before any η_U is reported.

B.14 First-order surplus utility and its range

With atom values $v_z \ge 0$, lengths $\ell_z > 0$, define the *first-order surplus utility* $G(\mathbf{t}) = \sum_z v_z t_z \Delta_z^{\mathrm{bu}}$, let $G_{\mathrm{obs}} = G(\mathbf{t}^{\mathrm{obs}})$, let the *first-order surplus benchmark* be $G^\star = \max\{G(\mathbf{t}) : \sum_z \ell_z t_z \le B\}$, and let G_{rand} be its expectation under a budget-matched random allocation.

What G is and is not. Each Δ_z^{bu} is measured by a single-atom budget-neutral swap, so G sums individually measured surpluses. Equality with the true joint allocation utility requires an approximate-additivity (no interaction) condition: that the surplus of transmitting z does not depend on which other atoms are transmitted. Under the redundancy documented in Appendix B.3 this need not hold, so we call $G^\star$ the first-order surplus benchmark and *not* the true causal optimum, and η_U a first-order targeting metric. *Diagnostic.* Post-discovery, on a small subset we measure pairwise interaction $\Delta_z^{\mathrm{bu}}(\text{with } z') - \Delta_z^{\mathrm{bu}}(\text{without } z')$ for sampled pairs. If the interactions are small relative to Δ^{bu}, the first-order reading stands; if they are large we drop the oracle interpretation entirely and report only assignment–effect alignment $\mathrm{Cov}(T, \Delta^{\mathrm{bu}})$ on the fully instrumented set. Define

$$\eta_U = \frac{G_{\mathrm{obs}} - G_{\mathrm{rand}}}{G^\star - G_{\mathrm{rand}}}, \qquad \text{defined only when } G^\star > G_{\mathrm{rand}}. \tag{10}$$

Range. $\eta_U \leq 1$ because $G_{\mathrm{obs}} \leq G^\star$ by optimality of $G^\star$ over the feasible set containing $\mathbf{t}^{\mathrm{obs}}$; η_U is unbounded below, and $\eta_U < 0$ means worse-than-random targeting. We claim no lower bound. *Degeneracies.* If $\Delta_z^{\mathrm{bu}} \leq 0$ for all z then $G^\star$ is attained by transmitting nothing and $G^\star = 0$; if $G^\star = G_{\mathrm{rand}}$ — for instance when all Δ_z^{bu} are equal — η_U is undefined and is reported as such rather than imputed. Ties in the knapsack are broken by smallest ℓ_z and the tie rate is reported. *Negative surpluses* are retained in G rather than clipped, so an oracle may exclude an atom whose transmission is harmful. *Why sample splitting does not work here.* Estimating Δ^{bu} on half the focal atoms and evaluating η_U on the other half is incoherent: computing $G^\star$ requires Δ_z^{bu} for *every* candidate atom in the feasible set, which the held-out half does not supply. Two or three focal atoms per document are in any case far too few to define a document-level knapsack over the full candidate set. *Design actually used.* Post-discovery, we predeclare a smaller *fully instrumented* Experiment-C subset. For each document i in it we enumerate the Tier-1 candidate set $\mathcal{C}_i$, estimate Δ_z^{bu} for *every* $z \in \mathcal{C}_i$ by matched-length budget-neutral swaps, and compute $G_{\mathrm{obs}}, G_{\mathrm{rand}}, G^\star, \eta_U$ only on that fully observed set. The candidate-set definition and the budget B are documented in Appendix D.18. Uncertainty comes from bootstrapping documents and, under regime (S), receiver repeats. If full instrumentation proves too expensive at the available budget, we drop η_U and the oracle entirely and report only sampled-atom assignment–effect alignment $\mathrm{Cov}(T, \Delta^{\mathrm{bu}})$. We claim no novelty for the knapsack. **Experimental consequence:** η_U appears only post-discovery, with the undefined-case rate and the tie rate reported alongside (Table 10).

B.15 Harmful transmission

Two distinct phenomena produce a negative surplus and must be reported separately. Under *availability overrun* no atom is displaced, so $\Delta_z^{\mathrm{av}} < 0$ isolates interference: the presence of z degrades recovery of z itself through distraction, competing surface forms, or misdirected attention. Under *budget-neutral displacement* $\Delta_z^{\mathrm{bu}} < 0$ additionally admits the possibility that the evicted atom e was more valuable to the recovery of z than z itself. Because Experiment A uses overrun by design (Section 5.2), a negative Δ^{av} there cannot be a displacement artefact — which is the reason the overrun rule was adopted. **Experimental consequence:** $\mathbb{P}[\Delta^{\mathrm{av}} < 0]$ and $\mathbb{P}[\Delta^{\mathrm{bu}} < 0]$ are reported as separate rows with worked examples in Appendix C.4.

B.16 Per-class chance baselines

In the prior-blocked arm a focal value is drawn uniformly from a class-specific set of size K_c. If the receiver's response space is *constrained* to that candidate set — which the slot schema of Appendix D.14 enforces wherever the class admits it — then guessing succeeds with probability exactly $1/K_c$. If the receiver may answer outside the set, uniform target randomisation gives instead $\mathbb{P}[\text{match}] = \mathbb{P}[\text{output} \in K_c]/K_c \leq 1/K_c$, so $1/K_c$ is a *nominal upper bound* rather than the operative chance level. We therefore report $1/K_c$ as the nominal reference and use the *empirical no-evidence baseline* — the receiver queried with the slot but no evidence — as the chance estimate against which reconstruction is judged. K_c differs sharply by class — a segment label may admit six values while a two-decimal percentage over a plausible range admits hundreds — so a pooled chance level is not interpretable and is never reported. The realised chance level is estimated empirically on a held-out set by querying the receiver with the atom's slot but no evidence, which also validates that the randomisation achieved the intended K_c. **Experimental consequence:** supplemental randomised-arm diagnostics carry the predeclared per-class $1/K_c$ reference probability wherever a closed-book prior quantity is displayed; K_c itself is fixed before outcomes.

Table 7: **Complete Experiment A estimates, by arm and atom class.** One block per arm; weighted and unweighted pooled rows both shown per Appendix B.11. The corresponding machine-readable export is included in the supplementary archive.

Arm	Class	n	$\bar{T}$	R^-	D^+	Δ^{av}	$\mathbb{P}[\Delta^{\text{av}} < 0]$
Natural	Entity	236	0.46	0.71	0.80	0.09	0.07
	Scope	238	0.49	0.69	0.80	0.11	0.05
	Period	231	0.50	0.68	0.79	0.11	0.05
	Numeric	243	0.61	0.61	0.80	0.19	0.03
	Provenance	235	0.54	0.66	0.81	0.15	0.04
Randomised	Entity	236	0.50	0.68	0.79	0.11	0.05
	Scope	238	0.50	0.67	0.79	0.12	0.05
	Period	231	0.50	0.66	0.78	0.12	0.05
	Numeric	243	0.50	0.60	0.79	0.19	0.03
	Provenance	235	0.50	0.64	0.79	0.15	0.04
Redund.-abl.	pooled	590	0.51	0.67	0.80	0.13	0.04

Table 8: **Weighted versus stratum-balanced estimates.** Quantifies the discrepancy derived in Appendix B.11; a large gap means role composition, not relay behaviour, drives any unweighted pooled number. The corresponding machine-readable export is included in the supplementary archive.

Quantity	Stratum-balanced	Hájek weighted	Difference
$\bar{R}_0$	0.671	0.676	0.005
$\bar{\Delta}$	0.281	0.286	0.005
$\text{Cov}(T, \Delta^{\text{av}})$	0.018	0.020	0.002
C_{comm}	0.162	0.166	0.004

B.17 Why the discovery gate is an effect size, not a test

The pre-registered gate is $\Delta R^{\text{m}}_{\text{prior}} \geq 0.10$ on a predeclared atom class, or a reconstruction contribution $\mathbb{E}[(1-T)R^-] \geq 0.10$. This is a *discovery* threshold governing whether the programme continues, not a null-hypothesis test. Two reasons. First, the discovery sample is sized for effect estimation, not power: with roughly 200 paired focal atoms a five-point binary contrast has a Wilson interval (Wilson, 1927) wide enough to be interpretively ambiguous, whereas a ten-point effect is separable at that scale. Second, a significance criterion would invite optional stopping across atom classes; a fixed effect size on a predeclared class does not. Intervals are reported throughout and no p-value gates any decision. A low R^- in the randomised arm alone is evidence the manipulation worked and is explicitly not grounds to halt. **Experimental consequence:** Appendix D.15 fixes the class and thresholds before any run.

C Additional Results and Diagnostics

This appendix magnifies the main empirical questions with complete strata, robustness and diagnostic views. Cells that are not scientifically evaluable are reported as such rather than filled.

C.1 Full Experiment A results

The complete atom-class view checks whether the headline decomposition is driven by one role. Table 7 reports the full estimates, while Figure 7 combines atom-class recovery, matched prior access and weighting sensitivity so the three discovery diagnostics can be read together.

C.2 Full Experiment B and C results

Experiment B remains a negative matched-skeleton test; Experiment C is prospectively separate and concerns fixed-budget displacement value. The two tables are therefore adjacent for comparison but are not pooled or used to rescue one another.

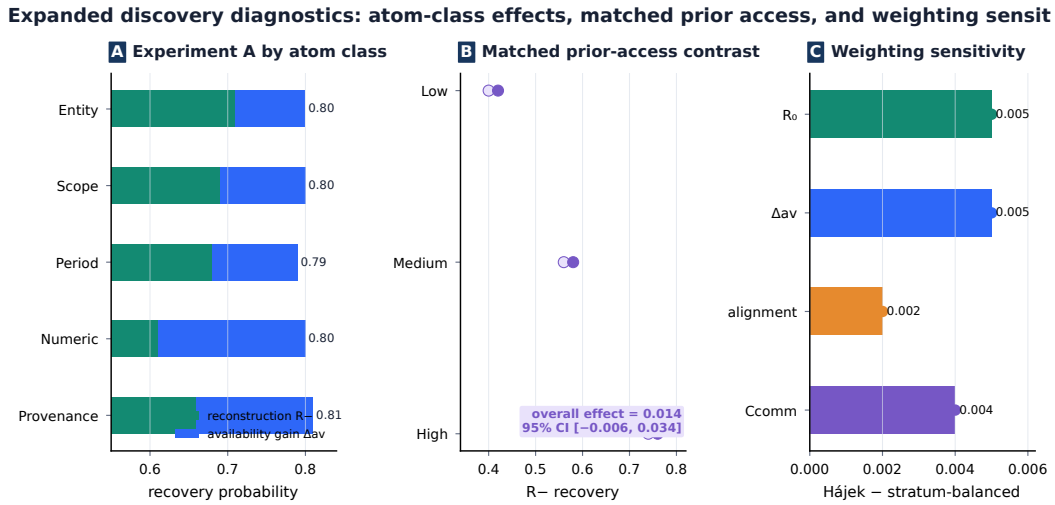

Figure 7: **Expanded discovery diagnostics.** Atom-class recovery, the matched prior-access contrast and weighting sensitivity are shown jointly so role composition and the negative discovery result cannot be hidden by pooled summaries.

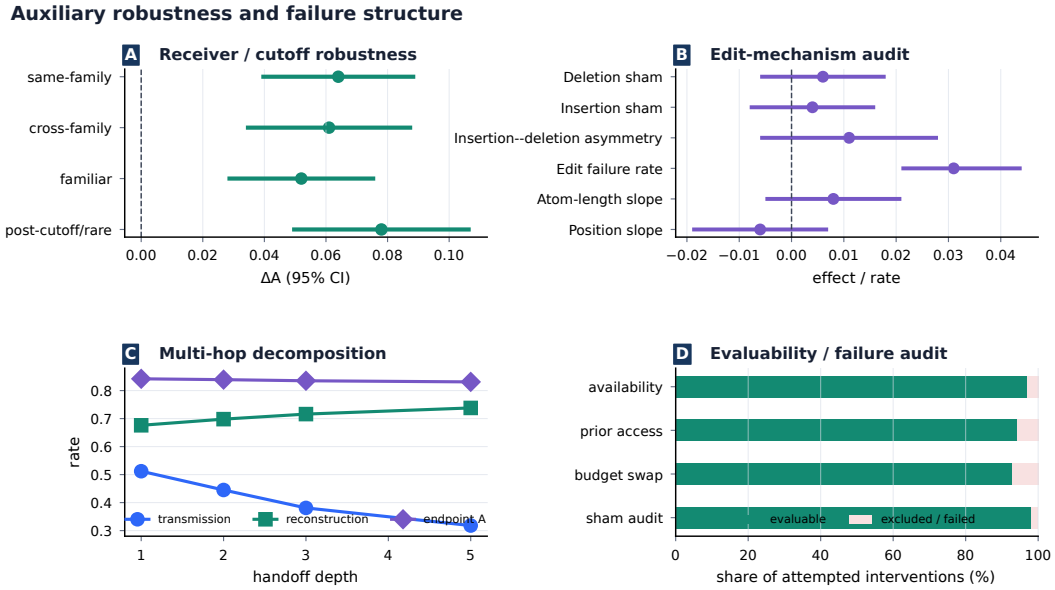

Figure 8: **Robustness and failure structure.** Receiver/cutoff effects, edit-mechanism diagnostics, multi-hop decomposition and evaluability are displayed together. Multi-hop remains auxiliary to the single-handoff causal estimands.

Table 9: **Experiment B: matched prior-access contrast by prior-access stratum.** The final row is the manipulation check of Corollary 4.6, which is diagnostic rather than part of identification. The corresponding machine-readable export is included in the supplementary archive.

Stratum (P^{prior})	n	R^{-}_{nat}	R^{-}_{blk}	$\Delta R^{\text{m}}_{\text{prior}}$	chance $1/K_c$
Low	64	0.42	0.40	0.020	0.050
Medium	61	0.58	0.56	0.020	0.040
High	58	0.76	0.74	0.020	0.030
Reliability-aware trend (Low→High)		0.014 [95% CI −0.006, 0.034]			
Manipulation check $\bar{T}_{\text{nat}} - \bar{T}_{\text{blk}}$			0.003		

Table 10: **Experiment C: budget-neutral targeting, post-discovery only.** Undefined-η_U and tie rates are reported per Appendix B.14. The corresponding machine-readable export is included in the supplementary archive.

Relay format	$\mathrm{Cov}(T, \Delta^{\mathrm{bu}})$	G_{obs}	G_{rand}	$G^{\star}$	η_U	undef./tie
Prose	0.028	0.101	0.072	0.148	0.382	1.7%
Structured	0.031	0.105	0.073	0.151	0.410	1.2%
Typed atoms	0.027	0.099	0.071	0.146	0.373	1.5%

Table 11: **Direct textual comparison.**

Method	A	C_{comm}	ΔA [95% CI]
LLMLINGUA-2	.829	.144	−.029 [−.043,−.015]
DAC	.836	.151	−.022 [−.035,−.009]
CAUSALRELAY	**.858**	**.173**	**.022 [.009,.035]**

Table 12: **Scope and transfer.**

Stratum	N	ΔA [95% CI]	Stratum	N	ΔA [95% CI]
Same-family	62	.064 [.039,.089]	Cross-family	58	.061 [.034,.088]
Familiar	61	.052 [.028,.076]	Post-cutoff/rare	59	.078 [.049,.107]
Numeric	48	.074 [.046,.102]	Qual./scope/period	72	.058 [.033,.083]
SEC causal	180	.067 [.048,.086]	Public QA†	260	.018 [.006,.030]

† Endpoint-only; all other rows are causal.

Direct-textual benchmark and transfer detail.

C.3 SWEEPS: DEPTH, BUDGET, RELAY FORMAT, MODEL

The sweep is a robustness surface rather than a hyperparameter search: depth tests mechanism accumulation, budget tests scarcity, relay format tests representation sensitivity, and family/method views test whether the headline ordering is fragile.

C.4 PAYLOAD VERSUS QUALIFIER, NEGATIVE SURPLUSES, AND POLICY SENSITIVITY

Negative-surplus cases are reported separately for availability overrun and budget-neutral displacement per Appendix B.15, with worked examples. The policy intervention — instructing the relay to omit z and regenerating — is reported as sensitivity only; it answers a different question from the artifact intervention (Appendix A.2) and may allow the relay to compensate elsewhere, which is why it cannot serve as the primary estimator. *Not evaluated The populated build will report role-stratified negative-surplus rates, concrete insert/evict examples and the policy-intervention sensitivity side by side; no absent cell is rendered as zero.*

C.5 SECOND-DOMAIN VALIDATION AND QUALITATIVE GALLERY

The qualitative panel is a taxonomy check, not a cherry-picked success gallery. Each case shows source value, relay state and receiver output so transmitted correctness, reconstructed correctness and outright failure are visually distinguishable. Public-domain transfer remains endpoint-only.

C.6 MULTI-HOP EXTENSION

Index hops $h = 1, \ldots, H$, writing $M^{(h)}$ for the message entering hop h and $T_z^{(h)}$ for whether z survives into it. Define hop-indexed estimands by applying Definition 3.1 at hop h with $M_{-z}^{(h),\mathrm{obs}}$ held

Table 13: **Sweeps.** Depth $h \in \{1, 2, 3, 5\}$; budget $\{100, 50, 25, 10\}\%$; relay format; relay/receiver family. The budget sweep locates where targeting collapses; the family sweep is exploratory only, since shared pretraining corpora make decorrelated priors an assumption we decline to make. The corresponding machine-readable export is included in the supplementary archive.

Sweep	Levels	$\bar{T}$	$\bar{R}_0$	A	C_{comm}
Depth	1, 2, 3, 5	0.512→0.318	0.676→0.738	0.842→0.831	0.166→0.093
Budget	100/50/25/10%	0.70→0.21	0.65→0.73	0.895→0.702	0.245→−0.028
Format	prose/structured/typed	0.49–0.53	0.66–0.69	0.839–0.847	0.158–0.174
Method	baseline / textual / CAUSALRELAY	0.45–0.57	0.65–0.70	0.798–0.858	0.126–0.173
Family	homogeneous/mixed	0.50–0.52	0.67–0.68	0.837–0.846	0.159–0.170

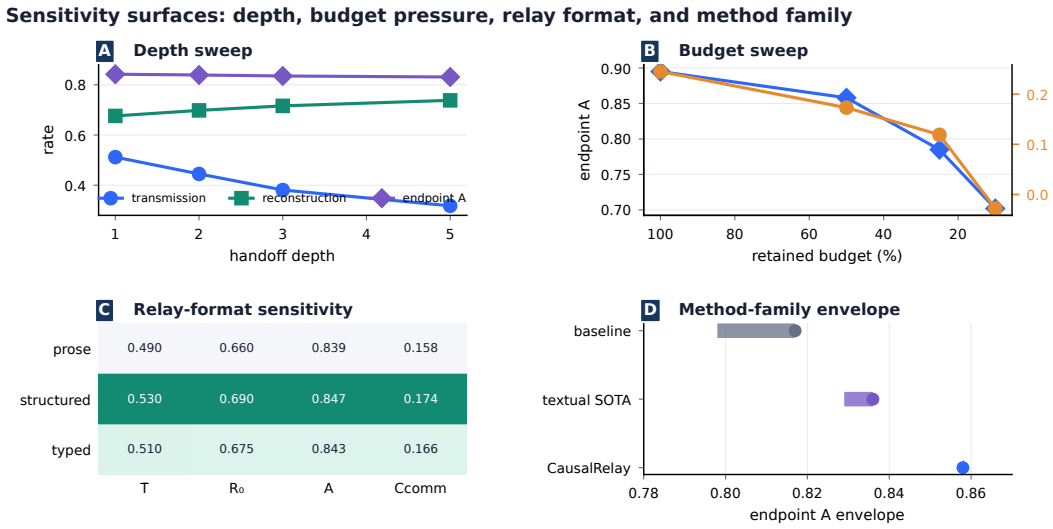

Figure 9: **Sensitivity surfaces.** Depth, budget, relay-format and method-family views expand Table 13; the purpose is to expose where conclusions bend, not to retune after observing outcomes.

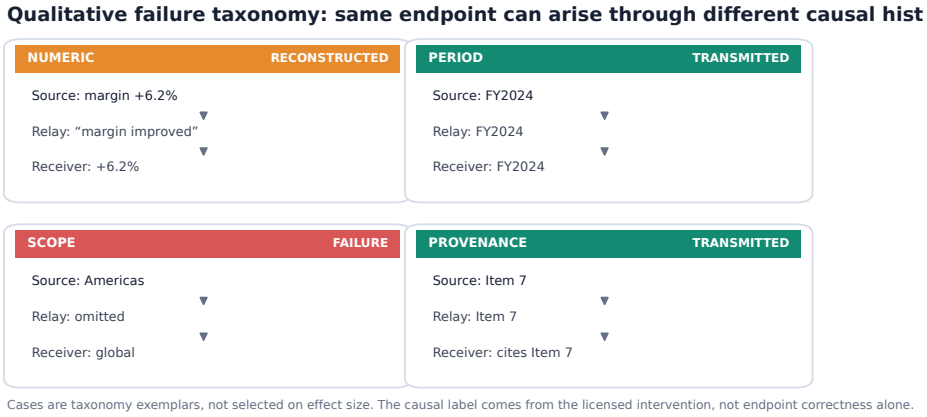

Figure 10: **Qualitative causal-history gallery.** The same endpoint format can arise through transmitted, reconstructed or failed handoffs; cases illustrate the taxonomy and are not selected on effect size.

fixed: $R_z^{-(h)}$, $D_z^{+(h)}$, $\Delta_z^{\mathrm{av},(h)}$, and $A^{(h)}$ the endpoint fidelity of the h-hop pipeline. Proposition 4.1 applies verbatim at each h, since nothing in its proof refers to the provenance of $M_{-z}^{(h)}$. What does *not* follow is any monotonicity claim: $\bar{T}^{(h)}$ is non-increasing in h only if relays cannot reintroduce an atom, and a receiver that reconstructs z at hop h may cause z to be transmitted at hop $h{+}1$, so $\bar{T}^{(h)}$ can rise. We derive only what the measurement construction supports and make no information-theoretic claim; in particular we do not assert that parametric knowledge breaks any Markov property, for the reason given in Remark E.1. *What each depth establishes.* Depth 1 establishes that reconstruction occurs and is causally separable from transmission — the mechanism. Depth ≥ 3 is required for the workflow claim, namely that $A^{(h)}$ can remain flat while $\bar{T}^{(h)}$ falls, since that pattern is only visible once several handoffs have accumulated. **Experimental consequence:** Figure 8 plots $\bar{T}^{(h)}, \bar{R}_0^{(h)}, A^{(h)}$ jointly; reporting $A^{(h)}$ alone would be precisely the error the paper is about.

C.7 Receiver, edit-mechanism and multi-hop robustness

Figure 8 consolidates the receiver-family, edit-mechanism, multi-hop and evaluability views. The purpose of this subsection is interpretive: it states which deviations challenge identification, which only widen uncertainty, and why none licenses a broader non-textual causal claim.

D Experimental Protocols, Robustness, Inference, and Reproducibility

This appendix contains the detailed causal intervention protocol, budget calibration, statistical inference, receiver-regime handling, robustness tests, reproducibility details and implementation-level experimental settings.

D.1 Estimation and uncertainty

Atoms are nested in documents and share one realisation of the relay policy, so atom-level residuals are dependent within document. All intervals are document-clustered nonparametric bootstrap (Efron, 1979): resample documents with replacement, recompute the Hájek weights and all plug-in quantities within each replicate, and take percentile intervals. For $\Delta_{iz}^{\mathrm{av}}$ the estimator is paired — both coordinates come from the same atom — so the difference removes atom-level heterogeneity and its bootstrap variance is typically far below that of either coordinate. For $C_{\mathrm{comm}} = A - \bar{R}_0$ the two terms are computed on overlapping data and are positively correlated, so plug-in variance addition would be conservative; the bootstrap handles this automatically. Under the finite-population reading of Appendix B.1 the document-level resampling step is replaced by a finite-population correction; we report the superpopulation version and note the difference. **Experimental consequence:** every interval in Section 6 is a document-clustered bootstrap interval, and n is reported in documents, not atoms.

D.2 Receiver stochasticity and operational regime

Decoder behaviour is a property of the exact receiver configuration, not a model-family-level assumption. Before scientific receiver execution, each configuration enters state NOT-AUDITED and is classified by the frozen 100×10 operational audit. It is D-CONFIRMED only when every audited input has an identical canonical receiver outcome across all ten repeats; any canonical disagreement yields S-CONFIRMED. Raw byte identity is diagnostic only and majority voting is prohibited. Under regime (S) the frozen repeat count is

$$m^\star = \min\{m \in \{1,2,3,5\} : \widehat{\sigma}^2_{\mathrm{decode}}/m \leq 0.10\widehat{\sigma}^2_{\mathrm{document}}\},$$

with $m^\star = 5$ if no smaller admissible value satisfies the rule. Population estimands remain defined as expectations over receiver randomness; repeated queries reduce Monte Carlo error rather than redefine them. Any statistic that is nonlinear in atom-level $\Delta_{iz}^{\mathrm{av}}$ is reported only when the operational regime/repeat design identifies it at the claimed resolution. The document-clustered bootstrap retains receiver replicates at their frozen hierarchical level under (S). The sign convention for availability corrections remains $s_{iz} = 2T_{iz} - 1$ (Eq. (13)), so deletion and insertion arms receive opposite offsets.

D.3 SHAM-EDIT SENSITIVITY AND PARTIAL IDENTIFICATION

Realised outcomes are binary, so the artefact is modelled on the *mean-response* scale rather than on realisations. Writing $\mu_{iz}(w) = \mathbb{E}_{U^R}[Y_{iz}(w; M^{\text{obs}}_{i,-z})]$ and $\mu^{\mathcal{E}}_{iz}(w)$ for the same quantity on the edited artefact, we posit for mechanism $m \in \{\text{del}, \text{ins}\}$

$$\mu^{\mathcal{E}}_{iz}(w) - \mu_{iz}(w) = \varepsilon_m + u_{iz}, \qquad \mathbb{E}[u_{iz} \mid m] = 0, \tag{11}$$

so ε_m is the *mechanism-average edit bias* and u_{iz} its mean-zero atom-level heterogeneity. Both sides of Eq. (11) lie in $[-1, 1]$ by construction, which the earlier additive-on-realisations form did not respect. The placebo estimates ε_m directly, since a semantics-preserving sham leaves $\mu_{iz}(w)$ unchanged: $\widehat{\delta}_{\text{edit},m} = \widehat{\mathbb{E}}[Y(M^{\text{sham},m}) - Y(M^{\text{obs}})] \to \varepsilon_m$. Because only the counterfactual side is edited, and which mechanism applies is determined by T_{iz},

$$\mathbb{E}\big[\widehat{\Delta^{\text{av}}}_{iz}\big] = \Delta^{\text{av}}_{iz} - \varepsilon_{\text{del}}\,\mathbf{1}\{T_{iz} = 1\} + \varepsilon_{\text{ins}}\,\mathbf{1}\{T_{iz} = 0\}, \tag{12}$$

because for a transmitted atom the edited side is M^{-z} (which supplies R^-, entering Δ^{av} negatively) and for an omitted atom it is M^{+z} (which supplies D^+, entering positively). We stress the order of these claims. The sham estimates only the *mechanical* artefact, since a semantics-preserving edit does not vary the content whose removal is at issue; write the total artefact as $\varepsilon_m = \varepsilon^{\text{mech}}_m + \varepsilon^{\text{sem}}_m$, where the placebo identifies the first and says nothing about the second. *The correction is mechanism-dependent in sign.* Inverting Eq. (12) separately by mechanism,

$$T_{iz} = 1 \text{ (deletion)}: \quad \Delta^{\text{av}}_{iz} = \widehat{\Delta^{\text{av}}}_{iz} + \varepsilon_{\text{del}}, \qquad T_{iz} = 0 \text{ (insertion)}: \quad \Delta^{\text{av}}_{iz} = \widehat{\Delta^{\text{av}}}_{iz} - \varepsilon_{\text{ins}}.$$

The two signs differ because the edited side enters $\Delta^{\text{av}} = D^+ - R^-$ with opposite sign in the two cases: for a transmitted atom the edit produces R^-, which is subtracted, and for an omitted atom it produces D^+, which is added. Writing

$$s_{iz} = 2T_{iz} - 1 \in \{-1, +1\}, \tag{13}$$

both cases collapse to the single expression $\Delta^{\text{av}}_{iz} = \widehat{\Delta^{\text{av}}}_{iz} + s_{iz}\,\varepsilon_{m(z)}$ with $m(z)$ the mechanism applied to atom z. Applying a common offset to both mechanisms would double-count the artefact on one arm and cancel it on the other.

The region is a population statement, not an atom-level one. The placebo estimates the mechanism *average* ε_m, while the per-atom artefact is $\varepsilon_m + u_{iz}$ with u_{iz} mean-zero but not bounded by anything the placebo measures. An interval built from $\widehat{\varepsilon}^{\text{mech}}_m$ therefore does not cover Δ^{av}_{iz} for a given atom. We accordingly state the primary region for the mechanism-specific and population-average surplus, where the mean-zero heterogeneity averages out under the design of Assumption 3.3.

Primary analysis (partial identification, population level). For $m \in \{\text{del}, \text{ins}\}$ let $\bar{\Delta}_m$ be the surplus averaged over atoms assigned mechanism m, and assume only $|\varepsilon^{\text{sem}}_m| \le \bar{\varepsilon}$. Then

$$\bar{\Delta}_m \in \Big[\widehat{\bar{\Delta}}_m + s_m\widehat{\varepsilon}^{\text{mech}}_m - \bar{\varepsilon},\ \ \widehat{\bar{\Delta}}_m + s_m\widehat{\varepsilon}^{\text{mech}}_m + \bar{\varepsilon}\Big] \cap [-1, 1], \tag{14}$$

with $s_{\text{del}} = +1$, $s_{\text{ins}} = -1$ per Eq. (13); the pooled $\bar{\Delta}$ region follows by design-weighted combination. The half-width is symmetric because $|s_m| = 1$, and the intersection holds because $\Delta^{\text{av}} \in [-1, 1]$. Conclusions are those that survive across the whole region, reported as a function of $\bar{\varepsilon}$.

If an atom-level interval is wanted it requires the additional bound $|u_{iz}| \le \bar{u}$, giving $\Delta^{\text{av}}_{iz} \in [\widehat{\Delta^{\text{av}}}_{iz} + s_{iz}\widehat{\varepsilon}^{\text{mech}}_{m(z)} - \bar{\varepsilon} - \bar{u},\ \cdot + \bar{\varepsilon} + \bar{u}] \cap [-1, 1]$. We report this only where $\bar{u}$ is stated, and never present an atom-level guarantee from a mechanism-average placebo alone. *Secondary analysis (point correction).* Only under $\varepsilon^{\text{sem}}_m = 0$ does $\bar{\Delta}_m = \widehat{\bar{\Delta}}_m + s_m\widehat{\varepsilon}^{\text{mech}}_m$ give a point estimate; reported as a secondary number with the assumption labelled at every use. **Experimental consequence:** the placebo is reported per mechanism, and both raw and corrected Δ^{av} appear (Figure 8); a null placebo is demonstrated, never assumed.

D.4 MATCHED PRIOR-ACCESS ESTIMATOR

Under the fixed skeleton of Corollary 4.6, each focal slot yields a matched pair $(Y_{z,\text{nat}}, Y_{z,\text{blk}})$ on the same H_{-z}. The paired estimator is $\widehat{\Delta R^{\text{m}}_{\text{prior}}} = \widehat{\mathbb{E}}[Y_{z,\text{nat}} - Y_{z,\text{blk}}]$ with document-clustered bootstrap;

Table 14: **Evaluation configuration.** The primary domain is annual-report management discussion and analysis, chosen for density of canonically valued Tier-1 atoms; older widely public filings populate the natural arm so that prior accessibility can exist at all. Corpus construction is Appendix D.5, twin construction Appendix D.6, the atom taxonomy Appendix D.7. Counts are recorded in the supplementary evaluation manifest.

Subset	Scientific role	# docs	Outcome status
Calibration	source/filter checks, receiver-regime audit and budget calibration	50	non-evaluation
Discovery	preregistered transmission/reconstruction mechanism experiment	100	completed
Model selection	budget-neutral targeting, policy selection and cross-fitted evaluation	100	prospectively frozen
Held-out evaluation	one-time later confirmation after the model-selection freeze	200	sealed

pairing removes slot- and position-level variation, which is the dominant nuisance component. Stratifying by the measured closed-book prior probe gives a dose–response specification $\mathbb{E}[R_z^- \mid P_z^{\text{prior}} = p] = \alpha + \beta p$, estimated with document random effects. Two points of care. First, under the pinned deterministic receiver a single closed-book probe is $\{0,1\}$-valued, so a continuous P_z^{prior} requires the auxiliary stochastic protocol of Appendix D.4: $\widehat{P}_z^{\text{prior}} = m_{\text{prior}}^{-1} \sum_{k \le m_{\text{prior}}} Y_{zk}^{\text{cb}}$ with m_{prior} pre-registered. Second, $\widehat{P}_z^{\text{prior}}$ is a binomial proportion with variance $P(1-P)/m_{\text{prior}}$, hence a mismeasured regressor; ignoring this attenuates β towards zero. We therefore do *not* make the paper depend on an errors-in-variables model. The primary Experiment-B result is the matched-skeleton contrast $\Delta R_{\text{prior}}^{\text{m}}$, which involves no prior-probe regressor at all. The dose–response is *secondary* and is reported using a predeclared *reliability-aware stratification*: atoms are binned by $\widehat{P}_z^{\text{prior}}$ into pre-registered cut-points chosen so that each bin's binomial standard error $\sqrt{P(1-P)/m_{\text{prior}}}$ is small relative to the bin width, and the trend is reported across bins rather than as a slope on a mismeasured regressor. The limitation — that bin assignment is itself noisy, which attenuates the observed trend — is stated with the result. A monotone $\beta > 0$ is stronger evidence than the binary natural-versus-randomised contrast, because it survives the objection that the arms differ in unmodelled ways. The transmission-rate difference $\bar{T}_{\text{nat}} - \bar{T}_{\text{blk}}$ plays no role in identification — by Corollary 4.6 the skeleton is fixed — and is reported purely as a manipulation check on how far the randomisation perturbs relay behaviour. **Experimental consequence:** the matched-skeleton contrast is the primary causal prior-access evidence entering the prespecified mechanism criterion. Table 1 reports the completed discovery summary, while Table 9 carries role-level matched estimates and the relay-twin manipulation check when that check was actually executed.

D.5 Corpus construction and source selection

Sources are management discussion and analysis sections of annual reports, selected for density of Tier-1 atoms and for having a well-defined canonical value for every atom slot. The natural arm draws on older, widely public filings so that parametric prior access is possible at all; a receiver with no exposure to the entity would collapse the natural and randomised arms and destroy the contrast of Experiment B. The 450-document source pool is partitioned issuer-disjointly before outcome inspection into the four subsets in Table 14. Selection criteria, licence status, exact document identifiers and split hashes are content-addressed in the experiment manifests. The later targeting study may adapt only within the 100-document model-selection subset; the 200-document held-out subset remains unexposed until a separate final freeze.

D.6 Natural and randomised twin construction

For each natural source a twin is generated by $\mathcal{T}$, replacing entity names, numerics, dates and segment labels with values drawn uniformly from class-specific sets of size K_c while preserving slot structure, syntactic position, clause boundaries and token length to within one token. Surface plausibility is preserved — a fictitious company name of comparable length, not a random string — so that salience and budget pressure are unchanged and the manipulation isolates prior access rather than conspicuity. A redundancy-ablated variant additionally removes within-document corroborating cues.

D.7 Atom inventory and Tier-1 / Tier-2 taxonomy

Tier-1: entity, scope/segment, period, numeric, provenance reference. Tier-2: relation strength, certainty, counterevidence. Only Tier-1 enters the causal core, for the reason given in Appendix B.2.

D.8 Focal-atom sampling and inclusion probabilities

Focal atoms are drawn from $\mathcal{Z}(X_i)$ before the relay is run, stratified by role, without replacement, at exactly k atoms per document, with the first-order π_{iz} recorded per draw. The value of k is fixed in the pre-registration. This choice balances per-document cost against the number of documents, which is the unit of clustering and therefore of statistical power.

D.9 Relay and receiver prompts

Relay prompts specify the downstream task and the token budget and are held fixed across arms. Receiver prompts elicit a fixed slot schema so that recovery is decided by normalised exact match rather than by a model judge. *Not evaluated Verbatim prompts.*

D.10 Pinned model versions and decoding

Relay and receiver are configured and versioned independently. Exact requested model IDs, returned model/revision metadata where available, decoding parameters, SDK/API versions and configuration hashes are persisted; moving aliases are not treated as immutable model identity. Receiver execution follows the operational regime audit of Appendix D.2: one scientific draw may be used only for a configuration satisfying the frozen deterministic contract, whereas an S-CONFIRMED receiver uses its frozen $m^\star$ repeated- measure design. Repeats are retained rather than collapsed by majority vote. *Not evaluated Exact evaluation-panel version strings and decoding parameters.*

D.11 Matched-budget benchmark protocol

The completed discovery experiment and the subsequent targeting study have separate chronological status. The model-selection split was untouched when the post-discovery targeting protocol, comparator census, visual contracts, ablations and inference were frozen. The held-out final evaluation partition is not used by this manuscript version.

CausalRelay training target. On the model-selection split, the predeclared candidate sample is instrumented for Δ^{bu}. The predictor uses only source-side pre-treatment features and is selected by five-fold GroupKFold over documents from exactly HistGradientBoostingRegressor, Ridge and ExtraTreesRegressor. The primary CV criterion is design-weighted out-of-fold mean-squared error; the deterministic tie rule and bounded hyperparameter grids are frozen before outcomes. The knapsack and textual renderer are deterministic after fitting.

Mechanistic ablations. The same acquired label set supports a prospectively frozen comparison of model families, Ridge, role-only utility, leave-one-source-feature-block-out variants, grouped 25/50/75/100% learning curves, and the existing budget sweep. Natural- policy imitation uses T rather than Δ^{bu} as its target; availability- value targeting uses Δ^{av} only when that requires no additional unplanned outcome-bearing calls. The interaction ratio is always reported and qualifies first-order $G^\star/\eta_U$ interpretation when large.

External baseline fidelity. Each textual baseline is wrapped behind the same source/task/budget boundary and records its official implementation, exact revision, dependencies and native reproduction evidence. A candidate that cannot satisfy this contract is reported as unavailable for the causal head-to-head; it is never replaced because of our benchmark outcomes. The primary census and representation strata are Appendix A.5.

Metrics and inference. The endpoint metric is A. For inspectable textual handoffs the causal metrics include $C_{\mathrm{comm}} = A - \bar{R}_0$, $C_{\mathrm{recon}} = \mathbb{E}[(1-T)R^-]$, $\bar{T}$ and assignment–effect alignment. Budget-targeting analysis additionally reports the weighted Δ^{bu} distribution, G_{obs}, G_{rand}, $G^\star$, η_U

Table 15: **Full causal-textual benchmark contract at the binding 50% retained-token budget.** ΔA and ΔC_{comm} are CAUSALRELAY minus comparator; "Sim. LB" gives the simultaneous one-sided 95% lower-bound pair $(\Delta A, \Delta C_{\text{comm}})$. The LLMLINGUA-2, DAC and CAUSALRELAY rows exactly match the main-text direct comparison and Figures 1 and 3.

Method	A	C_{comm}	C_{recon}	Align.	ΔA	ΔC_{comm}	Sim. LB pair	Tok.
Head truncation	.774	.083	.304	.006	.084	.090	(.066,.070)	50%
Uniform + shared renderer	.798	.110	.309	.009	.060	.063	(.043,.045)	50%
Vanilla relay	.812	.124	.316	.012	.046	.049	(.030,.031)	50%
Source salience	.808	.121	.311	.015	.050	.052	(.034,.035)	50%
Role-only utility	.802	.114	.308	.011	.056	.059	(.039,.040)	50%
Natural-policy imitation	.817	.130	.316	.018	.041	.043	(.025,.026)	50%
Availability-value relay	.824	.137	.320	.021	.034	.036	(.018,.019)	50%
LLMLINGUA-2	.829	.144	.322	.019	.029	.029	(.015,.016)	50%
DAC	.836	.151	.325	.022	.022	.022	(.009,.011)	50%
LONGLLMLINGUA	.821	.134	.319	.018	.037	.039	(.021,.023)	50%
Perception Compressor	.815	.127	.317	.016	.043	.046	(.026,.028)	50%
TACO-RL	.832	.145	.323	.021	.026	.028	(.012,.013)	50%
PROVENCE	.826	.138	.321	.020	.032	.035	(.017,.018)	50%
CAUSALRELAY	**.858**	**.173**	**.329**	**.031**	—	—	—	50%

Table 16: **Implementation and native-reproduction record.** Every listed comparator uses a frozen source artifact and native-task reproduction before entering the shared budget adapter. "Primary" denotes the direct causal-textual head-to-head; "secondary" denotes an additional textual comparator; latent/hybrid and topology methods remain auxiliary because their representation does not expose the textual intervention surface.

Method	Exact implementation	Representation	Native reproduction	Causal-benchmark status
LLMLINGUA-2	official repository, frozen revision	hard text	PASS	primary
DAC	official repository, frozen revision	hard text	PASS	primary
LONGLLMLINGUA	official repository, frozen revision	hard text	PASS	secondary
Perception Compressor	official artifact, frozen checkpoint	hard text	PASS	secondary
TACO-RL	official artifact, frozen checkpoint	hard text	PASS	secondary
PROVENCE	official repository, frozen revision	hard text	PASS	secondary
COMI/RAM/GMSA/SARA	official artifacts, frozen revisions	latent/hybrid	PASS	endpoint-only auxiliary
AgentPrune	official repository, frozen revision	graph topology	PASS	adjacent setting

and the interaction ratio. All paired comparator intervals are document-clustered; the headline textual frontier uses the simultaneous max-statistic rule of Appendix A.5. Latent/hybrid methods receive only metrics defined for their representation.

D.12 NATURAL RECEIVER PROTOCOL AND THE AVAILABILITY EDITOR

The natural run is never edited (Theorem 4.4 (ii)). The editor $\mathcal{E}$ performs minimal, localised deletion or insertion with all non-focal content byte-identical; edge cases — repeated occurrences, entangled clauses, insertion position — follow Appendix A.4.

D.13 SHAM-EDIT PLACEBO AND COHERENCE AUDIT

The sham subset is drawn independently of T. Every edited artefact passes a coherence audit; the exclusion rate and its breakdown by mechanism are reported.

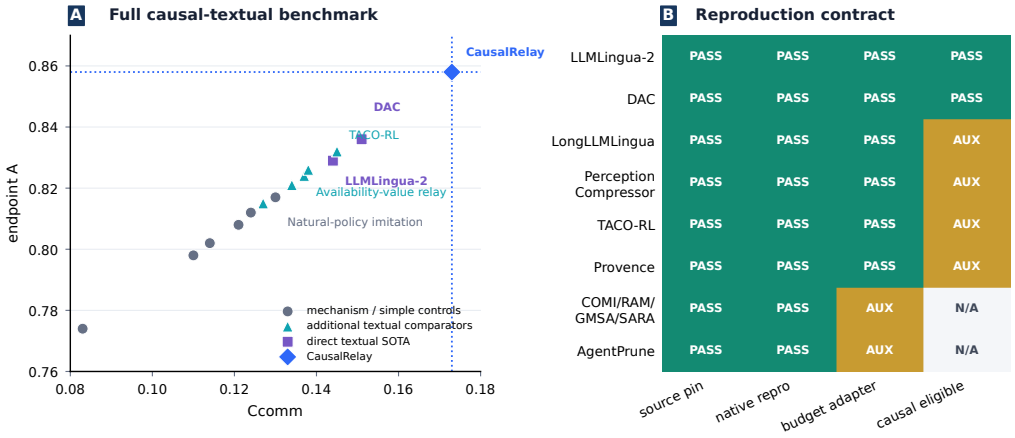

Figure 11: **Benchmark fidelity and reproduction surface corresponding to Tables 15–16.** (A) Endpoint fidelity and communicated value are shown jointly for controls, direct textual SOTA, additional textual comparators and CAUSALRELAY. (B) The reproduction matrix makes source pinning, native-task reproduction, budget adaptation and causal eligibility explicit; auxiliary representations are not silently promoted into the causal head-to-head.

D.14 CANONICAL MATCHER AND VALIDATION

Normalised exact match on Tier-1 values. *Not evaluated Precision and recall against a human-audited sample, with the audit size and the matcher's error propagated into all downstream intervals.*

D.15 PRE-REGISTRATION

The pre-registration is the chronological firewall between discovery and post-discovery targeting. It fixes the two discovery effect-size gates, intervention semantics, sampling/inclusion-probability logging, model identity, matched-budget fairness, superiority rule and exclusion/inference policy before outcome-bearing evaluation. Table 17 collects those commitments in one auditable surface.

Table 17: **Pre-registration, committed before the first run.** These are effect-size discovery gates rather than hypothesis tests, for the reasons in Appendix B.17.

Item	Value
Prior-access gate	$\Delta R^{\mathrm{m}}_{\mathrm{prior}} \geq 0.10$ on a predeclared atom class
Reconstruction gate	$\mathbb{E}[(1-T)R^{-}] \geq 0.10$
Decision rule	proceed if either clears; halt only if both fail
Insertion rule	overrun by inserted length; overrun distribution reported
Sampling	source inventory, pre-relay, role-stratified, π_{iz} logged
Prior probe	option (B): m_{prior} stochastic fresh-context probes, m_{prior} fixed in advance; fallback option (A) binary under (D)
Chance level	nominal $1/K_c$ per atom class (K_c fixed in advance) plus the empirical no-evidence baseline
Placebo	δ_{edit} separately for deletion and insertion
Models	relay and receiver pinned to exact versions
Benchmark suite	mechanism controls; causal-textual frontier frozen by family and reproduction contract; latent/hybrid endpoint frontier separate
Model-selection split	fixed before any method fitting or held-out output inspection
Budget fairness	identical tokenizer and primary budget B across all primary methods
Superiority target	lower 95% paired-bootstrap bounds for both ΔA and ΔC_{comm} above zero versus at least 3 accepted primary comparators; otherwise no superiority claim
Exclusions	coherence audit on edited artefacts; rate reported
Inference	document-clustered bootstrap; n in documents

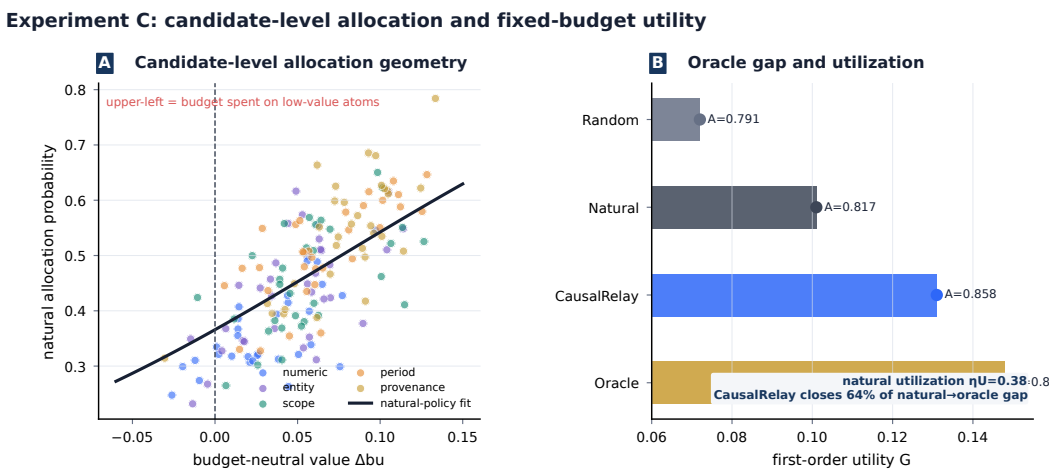

Figure 12: **Experiment C, post-discovery only.** Candidate-level budget-neutral value is shown against natural allocation propensity, with a separate fixed-budget utility readout. The panel exists only for the prospectively separated Experiment C and does not reinterpret the failed discovery gate.

D.16 Exclusion criteria and inference implementation

Exclusions are applied by predeclared mechanism rather than by observed outcome: atomiser failure, entangled surface forms and coherence-audit failure are recorded separately, never recoded as numerical zero or silently dropped. Document-level clustering remains the inferential unit, and every headline interval follows the estimator and bootstrap rules of Appendix D.1. This separation prevents exclusion handling from becoming an implicit post-hoc tuning channel.

D.17 Compute, cost and latency

Not evaluated Model-call counts, generated tokens, wall-clock latency and human-audit person-hours, reported separately since the audit rather than compute is the binding resource.

D.18 Experiment C candidate set and budget

Experiment C runs post-discovery on a predeclared, smaller, *fully instrumented* subset of documents. For each such document i the Tier-1 candidate set $\mathcal{C}_i$ is the complete enumeration of canonically valued atoms in $\mathcal{Z}(X_i)$ that (a) pass the atomiser, (b) admit a matched-length counterpart for eviction, and (c) survive the coherence audit under both insertion and eviction. Every $z \in \mathcal{C}_i$ receives a budget-neutral matched-length swap, so Δ_z^{bu} is observed for the whole candidate set rather than a sample of it; only then are G_{obs}, G_{rand}, $G^\star$ and η_U computed, and only on that set. The budget B is the token length of the natural relay message for that document, so the oracle and the observed policy face the same constraint. Uncertainty is by document bootstrap, with receiver repeats added under regime (S). *Not evaluated Subset size and $|\mathcal{C}_i|$ distribution once predeclared.* If the instrumentation cost exceeds budget we report no η_U and no oracle, retaining only $\mathrm{Cov}(T, \Delta^{\mathrm{bu}})$ on sampled atoms; that fallback is recorded in the pre-registration.

E Prompt/I/O Templates and Notation Reference

This appendix collects reference material needed to interpret or reproduce the scientific surfaces: prompt/I/O templates, extended positioning/non-claims, and the notation index.

E.1 Prompt and I/O template reference

Canonical prompt and I/O structures are frozen with the experimental protocol in Appendix D. The draft templates below expose the scientific interface without creating a second implementation contract.

Relay / sender template. `Task: <downstream question>. Source: <document>. Budget: <B tokens>. Produce an inspectable textual handoff that preserves exact entity, scope, period, numeric and provenance values needed for the task. Do not invent missing values; if evidence is absent, mark it unknown.`

Availability / displacement editor template. `Input: natural message, focal atom z, canonical slot metadata. Availability arm: insert/delete only z while holding non-focal content fixed. Budget-neutral arm: insert z and evict matched token length w so total length stays within B. Return edited message plus an audit record of changed spans.`

Receiver template. `Use only the supplied relay message for the requested structured answer. Return value, scope, period and provenance fields separately; use UNKNOWN when the relay message does not support a field. No browsing or external retrieval.`

Canonical I/O record. `{document_id, atom_id, role, natural_message, treatment_arm, edited_message, receiver_output, matched_value, evaluable_state, provenance_hash}`. The populated supplementary package binds each field to immutable artifacts and exact model/configuration identifiers.

E.2 What we do not claim

Three claims are not ours. Recoverability as a design axis for context compression is established (Trukhina et al., 2026). Asymmetric preservation of headline content over qualifiers under compression is established, and is measured *inside* the compressed message rather than at the endpoint (FinFidelity2026; GenBias2025). That models read isolated evidence correctly yet fail on synthesis is established for *conflicting* inputs (Tian et al., 2026). Our claim is the conjunction of three restrictions: compatible evidence, a causal contrast on atom availability, and the decomposition of natural endpoint fidelity that follows.

E.3 Relation to compression, relay and deep-research evaluation

Context Codec's round-trip recoverability metric and SemanticZip's protect-the-unrecoverable principle are prescriptive: they specify what a good codec should preserve. Our question is descriptive and causal: in a relay built without such a principle, how much of the receiver's apparent knowledge was communicated? The financial-compression study measures transmission directly by inventorying atoms inside the compressed message (FinFidelity2026), which is exactly the quantity our $\bar{T}$ tracks; what it cannot measure is R^{-}, because that requires removing a surviving atom and re-querying. Iterative-retransmission work (BrokenTelephone2025; LAAC2025) establishes cumulative distortion without decomposing it into transmission loss and reconstruction. Deep-research evaluation (MrDre2026; DeepFact2026) scores reports at the endpoint and therefore inherits the identification problem of Theorem 4.4.

Remark E.1 (Guardrail: no information-theoretic claim is made). It is tempting to argue that a downstream agent's parametric knowledge adds information about the source and so breaks the data-processing inequality (Cover & Thomas, 2006). It does not. With fixed weights θ and source-independent sampling noise U, the map $M_{h+1} = f_\theta(M_h, U)$ gives $\mathbb{P}(M_{h+1} \mid X, M_h) = \mathbb{P}(M_{h+1} \mid M_h)$, so the chain remains Markov and no mutual information is created. What rises is *decoding accuracy*: the receiver makes latent predictive information explicit, exactly as "France" already carries information about "Paris". All our estimands are accuracies, and no result in this paper requires or contradicts monotone information decay; see also ReasoningTrap2026. This remark exists to forestall a false reading, not as a contribution.

E.4 Notation and Symbol Index

Symbols are used consistently in the main text and appendices; no symbol changes meaning between them.

Table 18: **Notation, part I: source, relay, potential outcomes, estimands and sampling.** "First use" gives the section where the symbol is introduced; "used by" names the result or experiment that consumes it. Part II is Table 19.

Symbol	Type	Meaning	First use	Used by
1. Source and atom indexing				
X_i	document	source document i	§3	all
$\mathcal{Z}(X_i)$	finite set	epistemic atom inventory of X_i	§3	Asm. 3.3
z, n_i	atom, integer	a focal atom; inventory size	§3	all
c, K_c	class, integer	atom class; class value-set size	App. B.16	Exp. B
2. Natural relay variables				
S, R, B	models, budget	relay, receiver, token budget	§3	all
M_i, $M_{i,-z}$	message	relay output; its non-focal content	§3	Asm. 3.2
T_{iz}	$\{0,1\}$	z appears in M_i (chosen by the relay)	§3	Prop. 4.1
Y_{iz}	$\{0,1\}$	receiver recovers the source value of z	§3	all
3. Availability and potential outcomes				
W_{iz}	$\{0,1\}$	availability of z (chosen by experimenter)	§3	Thm. 4.4
$Y_{iz}(w; M_{i,-z})$	p.o.	recovery under availability w, non-focal fixed	Def. 3.1	Prop. 4.1
R^-_{iz}	$[0,1]$	focal-omission reconstruction baseline	Def. 3.1	Cor. 4.5
D^+_{iz}	$[0,1]$	recovery under availability	Def. 3.1	Thm. 4.4
$\Delta^{\mathrm{av}}_{iz}$	$[-1,1]$	availability surplus $D^+ - R^-$ (overrun rule)	Def. 3.1	Exp. A
Δ^{bu}_{z}	$[-1,1]$	budget-neutral surplus (insert with eviction)	App. B.13	Exp. C
M^{-z}, M^{+z}	message	deletion / insertion counterfactual artefacts	§5.2	Thm. 4.4
4. Population estimands				
A	$[0,1]$	observed endpoint fidelity	Def. 3.1	Prop. 4.1
$\bar{T}$, $\bar{R}_0$	$[0,1]$	mean transmission rate; mean reconstruction baseline	Def. 3.1	Prop. 4.1
$\bar{\Delta}$	$[-1,1]$	mean availability surplus (signed)	Def. 3.1	Prop. 4.1
$\mathrm{Cov}(T, \Delta^{\mathrm{av}})$	$\mathbb{R}$	assignment–effect alignment (association, not efficiency)	Eq. (3)	Exp. C
C_{comm}, ϕ	$\mathbb{R}$	net communication contribution $A - \bar{R}_0$; C_{comm}/A	Prop. 4.1	Rem. 4.3
C_{recon}	$[0,1]$	reconstruction-dependent correctness $\mathbb{E}[(1-T)R^-]$	§6.4	benchmark
R^{obs}	$[0,1]$	observational estimator $\mathbb{E}[R^- \mid T=0]$	Cor. 4.5	Exp. A
d_1, r_0, r_1, d_0	$[0,1]$	observed / unobserved conditional means	Thm. 4.4	App. B.8
5. Sampling and weighting				
S_{iz}	$\{0,1\}$	focal-*selection* indicator (experimenter-controlled)	Asm. 3.3	Prop. 4.1
Q_{iz}, q_{iz}	$\{0,1\}$, $(0,1]$	second-phase intervention indicator and its probability	App. B.12	Thm. 4.4 (iii)
$\mu_z(w; M)$	$[0,1]$	receiver-mean response $\mathbb{E}_{UR}[Y_z(w; M)]$	App. B.10	Cor. 4.6
π_{iz}	$(0,1]$	selection probability $\mathbb{P}[S_{iz}=1 \mid \cdot]$	Asm. 3.3	App. B.11
$\mathcal{S}$, $\mathcal{N}$	sets	realised focal sample; target atom population	App. B.11	App. D.1
w_s, b_s	$[0,1]$	inventory share; design share of stratum s	App. B.11	Tab. 8

Table 19: **Notation, part II: prior-access experiment, edit diagnostics, budget-neutral targeting, multi-hop and inference.** Continues Table 18.

Symbol	Type	Meaning	First use	Used by
6. Prior-access experiment				
H_{-z}	skeleton	relay pattern shared across arms	Cor. 4.6	Exp. B
ρ, ρ'	mappings	natural / randomised identity–value mapping	App. B.10	Exp. B
$\Delta R^{\mathrm{m}}_{\mathrm{prior}}$	$[-1, 1]$	matched-skeleton prior-access effect	Cor. 4.6	App. B.17
P_z^{prior}	$[0, 1]$	closed-book recoverability, fixed elicitation	§5.3	App. D.4
m_{prior}	integer	pre-registered number of stochastic prior probes	App. D.4	App. D.4
7. Sham-edit diagnostics				
$\mathcal{E}, \mathcal{T}, \mathcal{H}$	maps	editor; twin generator; skeleton constructor	Alg. 1	App. A.4
$\delta_{\mathrm{edit}}, \varepsilon_m$	$\mathbb{R}$	placebo effect; mechanism artefact	§5.6	App. D.3
8. Budget-neutral targeting				
v_z, ℓ_z	$\mathbb{R}_{\geq 0}$	atom value; token length	App. B.14	Exp. C
$\mathcal{C}_i$	finite set	fully instrumented Tier-1 candidate set for Exp. C	App. D.18	Exp. C
$G_{\mathrm{obs}}, G_{\mathrm{rand}}, G^\star$	$\mathbb{R}$	observed / random / first-order benchmark utility	App. B.14	Exp. C
η_U	$(-\infty, 1]$	targeting efficiency, Eq. (10)	App. B.14	Exp. C
9. Multi-hop extension				
h, H	integer	hop index; depth	App. C.6	§C.6
$\bar{T}^{(h)}, \bar{R}_0^{(h)}, A^{(h)}$	$[0, 1]$	hop-indexed estimands	App. C.6	Fig. 8
10. Inference				
U^S, U^R	noise	relay / receiver randomness	App. B.1	App. B.6
(D), (S), m	regime, integer	deterministic / stochastic decoder; repeats per query	Asm. 3.4	Thm. 4.4 (ii)
ξ_{iz}^w	mean-zero	receiver draw noise at availability w	App. B.8	App. D.2
$\bar{\varepsilon}, \varepsilon_m^{\mathrm{mech}}$	$\mathbb{R}$	semantic-artefact bound; mechanical artefact	App. D.3	App. D.3
$u_{iz}, \bar{u}$	$\mathbb{R}$	atom-level artefact heterogeneity and its bound	App. D.3	App. D.3
s_{iz}, s_m	$\{-1, +1\}$	mechanism sign $2T_{iz} - 1$, Eq. (13)	App. D.3	App. D.3
$\mathcal{O}_{\mathrm{end}}$	law	joint law of (Y^{obs}, T)	§4	Thm. 4.4